IET Computers & Digital Techniques

Research Article

IET Journals
The Institution of Engineering and Technology

Hybrid spin-CMOS stochastic spiking neuron for high-speed emulation of In vivo neuron dynamics

ISSN 1751-8601
Received on 30th June 2017
Revised 2nd January 2018
Accepted on 12th February 2018
doi: 10.1049/iet-cdt.2017.0145
www.ietdl.org

Steven D. Pyle[1] ✉, Kerem Y. Camsari[2], Ronald F. DeMara[1]

[1]*Electrical and Computer Engineering, University of Central Florida, 4000 Central Florida Boulevard, Orlando, FL, USA*
[2]*School of Electrical and Computer Engineering, Purdue University, 465 Northwestern Avenue, West Lafayette, IN, USA*
✉ *E-mail: steven.pyle@ucf.edu*

Abstract: The spintronic stochastic spiking neuron (S3N) developed herein realises biologically mimetic stochastic spiking characteristics observed within in vivo cortical neurons, while operating several orders of magnitude more rapidly and exhibiting a favourable energy profile. This work leverages a novel probabilistic spintronic switching element device that provides thermally-driven and current-controlled tunable stochasticity in a compact, low-energy, and high-speed package. In order to close the loop, the authors utilise a second-order complementary metal-oxide-semiconductor (CMOS) synapse with variable weight control that accumulates incoming spikes into second-order transient current signals, which resemble the excitatory post-synaptic potentials found in biological neurons, and can be used to drive post-synaptic S3Ns. Simulation program with integrated circuit emphasis (SPICE) simulation results indicate that the equivalent of 1 s of in vivo neuronal spiking characteristics can be generated on the order of nanoseconds, enabling the feasibility of extremely rapid emulation of in vivo neuronal behaviours for future statistical models of cortical information processing. Their results also indicate that the S3N can generate spikes on the order of ten picoseconds while dissipating only 0.6–9.6 μW, depending on the spiking rate. Additionally, they demonstrate that an S3N can implement perceptron functionality, such as AND-gate- and OR-gate-based logic processing, and provide future extensions of the work to more advanced stochastic neuromorphic architectures.

1 Introduction

Since the advent of the information revolution, research towards understanding and emulating brain-like intelligence by computer systems has yielded a plethora of neuronal models that underpin significant paradigms within artificial intelligence and machine learning. Such models range from computationally simple binary perceptrons [1], which output a one if the weighted sum is greater than a threshold, to complex biologically mimetic models, such as the Hodgkin–Huxley model that details somatic membrane potential dynamics due to ion channel conductance modulation [2]. Other intermediate models that can realise spike-based temporal information integration while being computationally efficient include the integrate-and-fire and leaky-integrate-and-fire models [3]. Such models have given rise to artificial neural networks (ANNs) that have performed tasks such as stock market prediction [4, 5] and defeating a 9-dan Go player [6].

Although the capability and utility of ANNs continue to increase, they are typically implemented on Von-Neumann style hardware that does not realise the ultra-low-power ultra-parallel capabilities of biological systems due to the Von-Neumann bandwidth bottleneck between the memory and processor [7]. Meanwhile, neuronal-inspired VLSI circuits and systems have been and are being researched and developed to take advantage of the inherent low-power and highly parallel nature of ANNs to achieve viable results [7]. However, much of the work towards hardware-based ANN designs have focused strictly on deterministic computational approaches, which lack the stochastic sub-threshold spiking characteristics found in in vivo neuronal measurements [8]. The stochastic dynamics of biological neural systems have been shown to be beneficial for signal detection, decision-making, memory recall, attention, Bayesian inference, and other neuronal behaviours [9–11]. It is widely acknowledged, stemming from the biological perspective, that in order to truly implement brain-like circuits and systems, some sufficient level of stochastic facilitation will be essential. Several recent works have aimed towards realising stochastic spiking neuronal hardware using both traditional CMOS as well as circuits utilising the intrinsically stochastic behaviours of some emerging devices. As shown in the following section, the proposed spintronic stochastic spiking neuron (S3N) circuit improves the previous designs by leveraging a novel spintronic probabilistic-bit (p-bit) device [12] that combines thermally driven stochasticity with electrically tunable bias to enable high-speed and low-power Poisson spike train generation within a compact design.

This paper is organised as follows: Section 2 details previous neural circuit designs utilising both CMOS and emerging devices. Section 3 describes the design and operation of the spintronic p-bit device utilised in the proposed design. Section 4 defines the S3N circuitry and the synapse circuit utilised herein. Section 5 delineates the simulation methodology and results. Section 6 provides a discussion of the limitations of the approach as well as its impact and extensions. Section 7 concludes the paper.

2 Previous works

Numerous works have aimed towards the realisation of neuromorphic circuits and systems, of which we review a recent selection of both deterministic and stochastic neural circuit approaches in this section.

2.1 Deterministic neural circuits

Neuromorphic circuits utilising CMOS technology have been developed ranging from low-power analogue designs [13–17] to fully digital designs that could be realised on field programmable gate srrays (FPGAs) [17, 18]. Additionally, VLSI neuromorphic chips have begun to emerge, demonstrating the feasibility of running various large-scale neural systems at low power on specialised hardware [7, 19].

Recent works have coalesced a large variety of specialised CMOS circuits capable of emulating a majority of the dynamics found in biological neurons, including temporal spike summation, spike-event generation, spike-frequency adaptation, thalamic relay, and dendritic tree effects [17]. Such circuits have been integrated to realise LIF models in both analogue and digital hardware, as previously discussed, as well as biophysically mimetic Hodgkin–Huxley models in mixed-signal VLSI [13].

Large-scale neuromorphic VLSI research projects such as IBM's TrueNorth and Stanford's Neurogrid have developed chips that demonstrate the impressive integration of up to millions of neurons and billions of synapses at exceptionally low levels of power consumption [7, 19]. While neurogrid focused on realising more biologically mimetic neuronal characteristics based on ion-channel modulation, TrueNorth abstracts neuronal spike dynamics into event-driven general purpose computations capable of implementing a variety of neuromorphic models. Some impressive capabilities of such hardware are the implementation of real-time multi-object recognition using just 63 mW [7] and semantic emotion recognition at just 50 μW [20]. Although the highly parallel nature of such systems has demonstrated impressive results at low power, the clock-rate is severely limited. For instance, TrueNorth operates at 1 kHz and Neurogrid at an average of 3.7 Hz compared to the GHz speeds of modern Von-Neumann processors.

In addition to CMOS-only neuromorphic circuits, hybrid designs utilising both CMOS devices and emerging devices, such as spintronics and memristors, have been developed. Memristor-based neural circuits have demonstrated a variety of behaviours found in cortical neurons such as regular spiking, intrinsic bursting, chattering, and fast spiking [21] within a compact one-memristor and three-transistor design, although the slow switching speed of memristors, compared to CMOS and spintronics, limits the operational speed of the circuit. Spintronic neural circuit designs have also been proposed that are of high speed (on the order of nanoseconds) and low area [22–24].

2.2 Stochastic neural circuits

Table 1 lists recent representative stochastic neural circuits and contrasts them to the proposed S3N design. For instance, IBM's TrueNorth chip, a purely digital CMOS design, has the capability of introducing stochasticity into neurons by stochastically varying the threshold voltage according to a pseudo-random number generator (PRNG), such as a linear feedback shift register [7]. Although PRNGs are effective for digital CMOS designs, they are not ideal because they lack true uncorrelated randomness and require significant area compared to designs using emerging devices. The 'spikes' of the TrueNorth chip are simply digital event signals that are synchronised and updated at a rate of 1 kHz.

Various hybrid CMOS/memristor designs have been proposed that utilise the stochastic formation of the conductive filament under weak programming conditions to determine if the circuit has 'spiked' or not [25–27]. This requires the circuit to cycle through a write phase, where the input signal (either voltage or current depending on the technology) attempts to switch the memristor, a read phase, where the circuit senses the state of the memristor and determines if a spike has occurred or not, and then a reset phase, where the memristor must be reset by a large deterministic reset signal prior to the next write cycle. This cycling behaviour as well as the long switching time of memristors limits the speed of the design, and the reset requirement after each read significantly increases the energy requirement.

Hybrid CMOS/phase-change neurons operate similarly to CMOS/memristor stochastic neurons by integrating the input signal into the crystallisation of atomic structure in chalcogenide phase-change materials, increasing the conductance [28]. Once the conductance reaches a certain threshold, a spike is emitted, and then a strong reset pulse returns the material to its amorphous phase, ready to begin the integration-crystallisation cycle again. The stochasticity of the device arises from the random atomic configuration in the amorphous state, which affects the rate of crystallisation in the integration phase.

Hybrid CMOS/spintronic circuits have been proposed to leverage the stochastic switching of low-energy-barrier nanoscale magnets in the thermally activated switching regime to realise stochastic neuron functionality in a write-rest-read-reset scheme similar to hybrid CMOS/memristor designs previously described [29–31]. Additionally, an asynchronous approach has been proposed to utilise ultra-low-energy barrier nanomagnets, which will stochastically switch due to thermal noise without the presence of magnetic or electrical bias, and can be tuned with such a bias [29]. Although the aforementioned approach is considered a spiking neuron design, it is not directly evident how distinct spike events are generated since just the state of the spintronic device is continuously sensed and propagated. As shown in the table, hybrid CMOS/spintronic designs are typically much faster and lower area than alternative stochastic neuron approaches and offers true uncorrelated randomness due to the utilisation of thermal noise. Thus, the circuit design proposed herein, as will be described more in depth in Section 4, takes a hybrid CMOS/spintronic approach to generate high-speed distinct asynchronous spike events at a rate tunable via input current by leveraging the thermally driven stochasticity of ultra-low-energy barrier nanomagnets.

3 Probabilistic spintronic logic device

The primary building block of the circuit proposed in this paper is a hardware unit called the probabilistic bit (p-bit) (Fig. 1*a*), which was recently described in the context of a novel type of probabilistic logic for Boolean and non-Boolean problems [12, 32]. The device combines a heavy metal (HM) exhibiting giant spin hall effect (GSHE) and a magnetic tunnel junction (MTJ) whose free layer magnetisation is modulated by the GSHE layer. Unlike standard experiments combining the HM with an MTJ that utilises ferromagnets with energy barriers of the order of 40–60 kT [33, 34], the p-bit uses an unstable ferromagnet, with an energy barrier of 0–1 kT, which can be obtained by either reducing the volume of a stable magnet [35] or by using circular magnets that effectively have no barrier in the absence of a geometrically preferred easy axis [36].

The MTJ resistance is modulated by the free layer magnetisation and an average resistance ($R_0 = \frac{1}{2}[G_P + G_{AP}]^{-1}$) is placed between the MTJ and two supply voltages (V^+) and (V^-), which creates a voltage divider circuit that produces a voltage fluctuating around the symmetry point of the CMOS inverter chain (Fig. 1*b*). In the absence of any GSHE current, the magnetisation fluctuates with average $\langle m_z \rangle = 0$ and the inverter chain amplifies this signal to produce rail-to-rail spikes between 0 and VDD. An input current into the GSHE generates a spin current that influences the magnetisation of the circular magnet, which effectively biases the spike probability.

The device components for modelling the GSHE/MTJ of the p-bit are represented by individually benchmarked circuit models. These models couple magnetisation dynamics to underlying transport equations [37, 38] that extend standard charge-based circuits to 'spin-circuits' by explicitly considering spin and charge voltages at each node, extending conventional resistors to 4×4 matrices. For example, the GSHE model converts a charge current

Table 1 Comparison of stochastic spiking neural circuits

	[7]	[25–27]	[28]	[29–31]	*Proposed herein*
Technology	CMOS	Hybrid CMOS/ memristor	Hybrid CMOS/phase change	Hybrid CMOS/ spintronic	Hybrid CMOS/ spintronic
Source of stochasticity	PRNG	Memristor switching probability	Random atomic configuration	Thermal energy	Thermal energy
spike implementation	event signal	write-read-reset cycle	integrate-fire-reset	write-rest-read-reset cycle, device sense	intrinsic circuit behaviour
spike time-scale order	1 ms	1–10 μs	10–100 ns	1 ns	10 ps
normalised device count	>10×	∼1×	∼2×	∼0.5 ×	1×

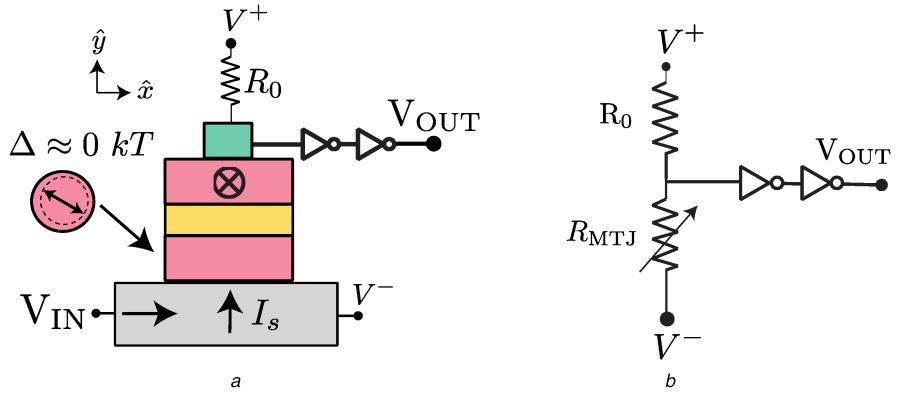

Fig. 1 *p-bit device schematic and equivalent READ circuit*

(a) The bottom layer represents an HM exhibiting the giant spin hall effect (GSHE) 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 ($E_B = 0$ kT) that fluctuates in the z–x plane in the presence of thermal noise. The MTJ is connected to an average resistance R_0 creating a fluctuating voltage that is amplified by two inverters, ***(b)*** Circuit equivalent READ circuit is also shown

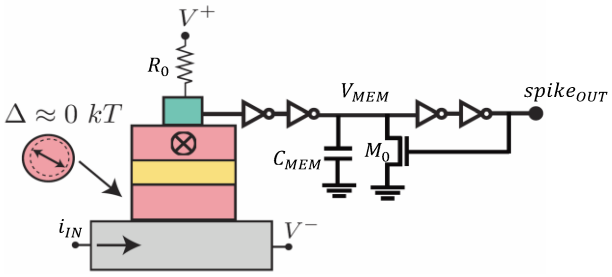

Fig. 2 *Spintronic stochastic spiking neuron circuit*

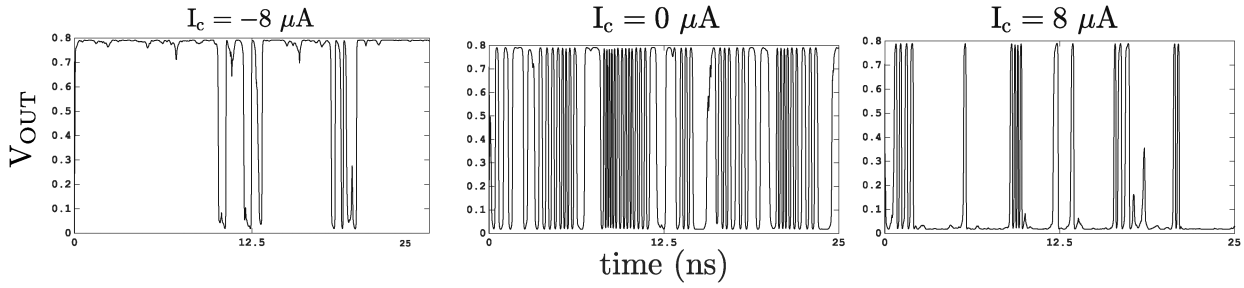

Fig. 3 *Transient response of the p-bit: The transient response of the p-bit at different input currents (I_c) flowing in the GSHE layer is shown. When there is no current, the output fluctuates between 0 and VDD = 0.8 V with equal probability, p = 0.5 and a mean of VDD/2. A positive input current in the GSHE increases the average output probability of 0 and a negative current increases the average output probability of VDD*

Table 2 P-bit simulation parameters

Parameter	Value
RA product (MTJ)	10 Ω-μm^2
Circular magnet diameter, thickness (FM)	50 nm, 0.5 nm
saturation magnetisation, ms	300 emu/cc
damping coefficient, α	0.01
HM length, width, thickness, spin-flip length, spin-hall angle, HM resistivity	50 nm, 50 nm, 3.15 nm, 2.1 nm, 0.5, 200 μΩ-cm
CMOS technology	PTM 14 nm-HP FinFET, VDD = 0.8 V
temperature	300 K
supply voltages	V + = VR + VDD/2 V− = VDD/2, VR = 0.25 V

to a spin-current that acts as a spin-torque input to the stochastic LLG module that solves for magnetisation dynamics, simulated self-consistently on a SPICE platform. The details of the spin-circuit formalism are given in [37]. For the CMOS inverters, predictive technology models (PTMs) are used [39]. Assuming short-circuit conditions where the FM absorbs all of the non-collinear spin-current, the GSHE layer produces a current I_S related to the charge current, running in the GSHE, I_c as

$$I_S = (I_C)\left(\frac{\theta L}{t}\right)\left(1 - \operatorname{sech}\left[\frac{t}{\lambda}\right]\right) \tag{1}$$

$$\begin{aligned}\frac{d\hat{\boldsymbol{m}}}{dt} &= -|\gamma|\hat{\boldsymbol{m}} \times \boldsymbol{H}_{\text{eff}} - \alpha|\gamma|\hat{\boldsymbol{m}} \times \hat{\boldsymbol{m}} \times \boldsymbol{H}_{\text{eff}} \\ &- \frac{1}{qN_i}\hat{\boldsymbol{m}} \times \hat{\boldsymbol{m}} \times \boldsymbol{I}_s + \frac{\alpha}{qN_i}\hat{\boldsymbol{m}} \times \boldsymbol{I}_s\end{aligned} \tag{2}$$

where θ is the spin-Hall angle, L, t, and λ are the length, thickness and the spin-flip length of the GSHE layer, respectively [38]. The spin-current flows normal to the plane of the magnet (y-direction in Fig. 2*a*) spatially, with a spin-polarisation that is in the $\pm$ z direction. A sufficiently large current can reduce the spiking probability or entirely suppress it by pinning the magnetisation of the circular magnet. Fig. 3 shows the transient V_{OUT} response as a function of different biases using the simulation parameters in Table 2.

The magnetisation dynamics of the circular magnet are obtained by solving the stochastic Landau–Lifshitz–Gilbert equation with the spin-torque component from the GSHE layer, and in the presence of the thermal noise within a monodomain approximation.

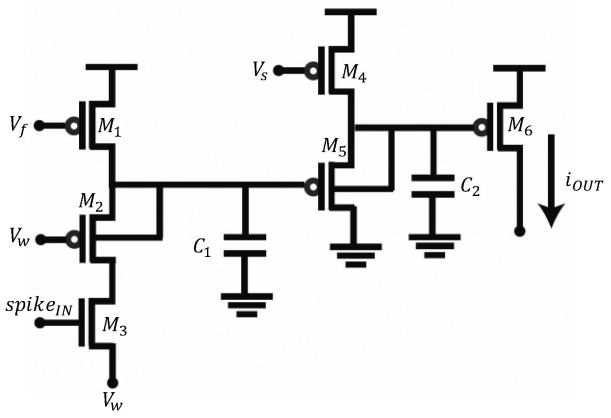

Fig. 4 *Neuromorphic synapse circuit used herein*

Table 3 S3N and synapse simulation parameters

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

We assume that the circular magnet does not possess any in-plane anisotropy ($H_{\mathrm{K}} = 0$) but has a large demagnetising field that keeps it in the plane. The effective field of the magnet is written as $\boldsymbol{H}_{\mathrm{eff}} = -4\pi M s\, m_y\, \hat{y}$. The thermal noise field is assumed to have three uncorrelated components with zero mean in all three directions ($i = x,y,z$) in cgs units: $H_i^{\mathrm{th}} = 0$, $\left(H^{\mathrm{th}}\right)_i^2 = (2\,\alpha\, kT/\gamma\, M_{\mathrm{s}}\ \mathrm{vol.}\ \Delta t\ )$, where α is the damping coefficient of the magnet, γ is the gyromagnetic ratio for the electron, M_{s} is the saturation magnetisation, and vol. is the volume of the magnet. The s-LLG equation is shown in (2) where q is the charge of an electron, N_i is the number of spins in the total volume, $N_i = M_{\mathrm{s}}$ vol/μ_{B}, where μ_{B} is the Bohr magneton, and $\boldsymbol{I}_s = |I_s|\ \hat{z}$ is the present coordinate system.

3.1 Equivalent read circuit

We assume a bias-independent MTJ conductance of the form $G_{\mathrm{MTJ}} = G_0\left(1 + P^2 m_z\right)$ [40]. The average resistance R_0 is chosen to be the average of G_{P} and G_{AP} so that when the average magnetisation is 0, the input to the inverter chain becomes halfway between V^+ and V^-. The supply voltages are adjusted such that the inverter characteristics are symmetric between $(\mathrm{V}^+ + \mathrm{V}^-)/2$. We assume that the inverter characteristics can be shifted accordingly by sizing the p- and n-type FETs as needed. It is important to note that the net voltage drop between V^+ and V^- can be small so that the READ voltage does not disturb the current state of the magnet through the fixed layer of the MTJ. This is possible in the presence of the inverter chain that provides gain and isolation.

4 Hybrid spin-CMOS stochastic spiking neuron

The S3N circuit developed herein is depicted in Fig. 2. 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 n-type metal-oxide-semiconductor (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, $\mathrm{spike}_{\mathrm{OUT}}$ will go high, subsequently turning on M_0, which discharges C_{MEM} and then sets $\mathrm{spike}_{\mathrm{OUT}}$ low. Thus, generating brief pulses, or spikes, at $\mathrm{spike}_{\mathrm{OUT}}$ 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 GSHE used herein. Therefore, alternative designs of p-bit devices that utilise alternative methods of magnetic bias, such as the magnetoelectric effect [41], can be readily implemented with the S3N scheme, providing future avenues of exploration and improvement to the design.

4.1 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 [42] and depicted in Fig. 4 is utilised 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 utilise 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 in Section 6, alternative synaptic architectures that prefer area efficiency over biological mimicry, such as crossbar arrays, could be utilised with the S3N for dense stochastic spiking neural network computational paradigms.

5 Results

The simulations provided in this section were performed in HSPICE using 14 nm FinFET transistor models [39]. The p-bit device was simulated using benchmarked SPICE spin-circuit models as described in Section 3. All inverter and synapse transistors are minimally sized with just a single fin, and M_0 has five 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}, 10 fF 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 3.

5.1 Stochastic spiking neuron

Fig. 5 shows the results of a single S3N neuron receiving a stepwise increasing input current. Every nanosecond, the input current is increased by 50 nA. $\widehat{m_z}$ is the z component unit vector of the magnetisation 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

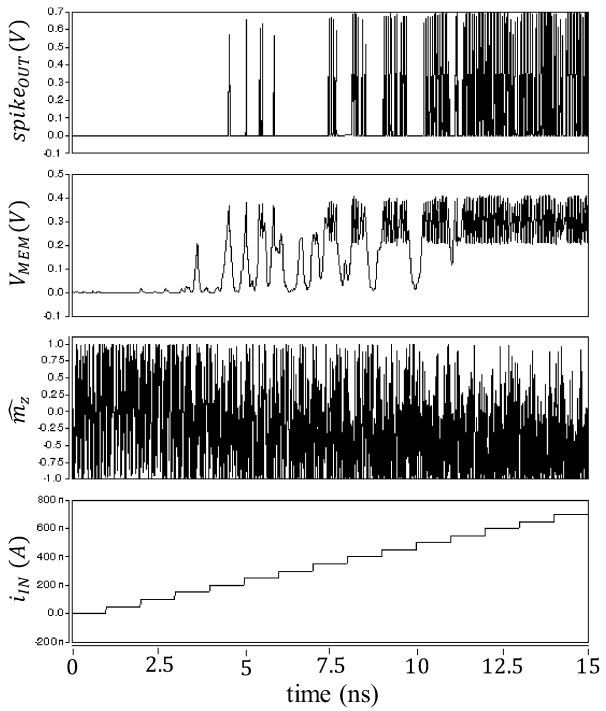

Fig. 5 *Stochastic spiking neuron simulation graphs illustrating, from the bottom up, i_{IN}, $\widehat{m_z}$, V_{MEM}, and $spike_{OUT}$*

for a long enough period, V_{MEM} will reach V_{th} and a spike is generated at $\mathrm{spike_{OUT}}$ 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 0.8 μA, which elicits a very high rate of spiking is 9.6 μW, and with an input current of 0 μA, which elicits almost no spiking, the SSN uses just 0.6 μW. The average spike width from (V_{dd}/2) to (V_{dd}/2) is just 15 ps. The average spike interval during high rates of spiking, such as with an input current of 0.8 μA, is about 120 ps.

5.2 Synaptic dynamics and weight control

The S3N combined with the second-order synapse circuit described in Section 4 was simulated by connecting the $\mathrm{spike_{OUT}}$ of the S3N circuit to the $\mathrm{spike_{IN}}$ terminal of the synapse and applying a fixed current of 0.5 μA at the i_{IN} terminal of the S3N with V_{w} set to 0.14 V, 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 Fig. 6*a*, 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 elicit 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}. Fig. 6*b* shows the effect of decreasing V_{w}, which effectively increases the weight of the synapse. A single-input S3N with an input current of 0.7 μA is used to generate the spike pattern $\mathrm{spike_1}$, which is then fed to the $\mathrm{spike_{IN}}$ terminal of a synapse, whose resulting output current is used as the input current for another S3N to generate the spike pattern $\mathrm{spike_2}$. 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 $\mathrm{spike_2}$ 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 0.14–0.2 V is enough to modulate the output from very high spiking to almost no spiking.

5.3 Boolean two-input perception

In order to demonstrate rudimentary computational capabilities utilising 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 whose 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 0.14 V, 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 Fig. 7*a*, where input current pulses of 0.7 μA are applied to input 1 from 2–4 ns and from 6–8 ns and to input 2 from 4–8 ns. As shown, a high rate of spiking activity at the output ($\mathrm{spike_3}$) occurs when either input is firing. Fig. 7*b* shows the results of the same simulation, but with V_{w} at 0.2 V, 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.

6 Discussion

In this section, the limitations of the proposed realisation are identified and discussed. Additionally, discussions are provided for extending this work to more advanced neuromorphic architectures that can utilise the stochastic nature of the proposed design.

6.1 Limitations of the proposed implementation

The primary contribution of this work 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 utilises voltage levels for weight-value implementation, which would require additional memory and programming elements in order to implement the

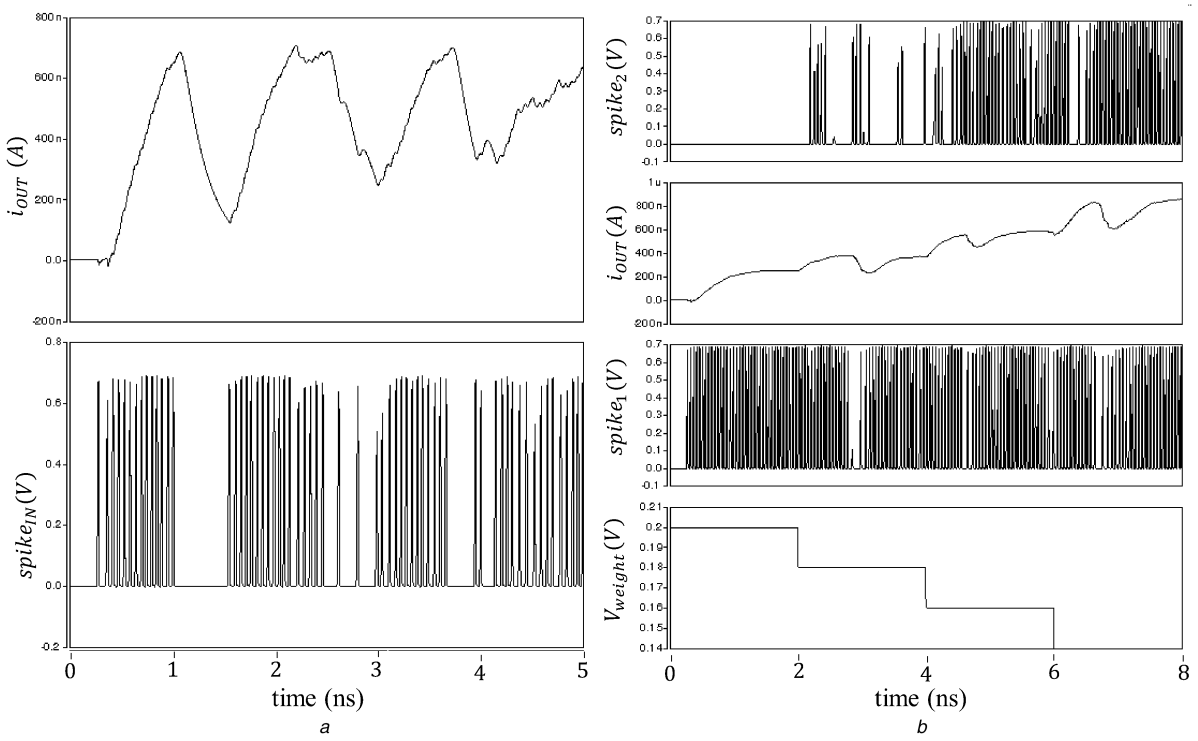

Fig. 6 *Simulation transients of stochastic spiking neuron with neuromorphic synapse*
(a) Neuromorphic synapse output current (top) in response to input spikes from S3N (bot), ***(b)*** From the bottom up, increasing the effective weight of the synapse by decreasing V_w, the output spike train from an S3N with an input current of 0.7 μA – which is used as the input to the synapse – the output current of the synapse, and the resulting spike train from driving another S3N with the synaptic current

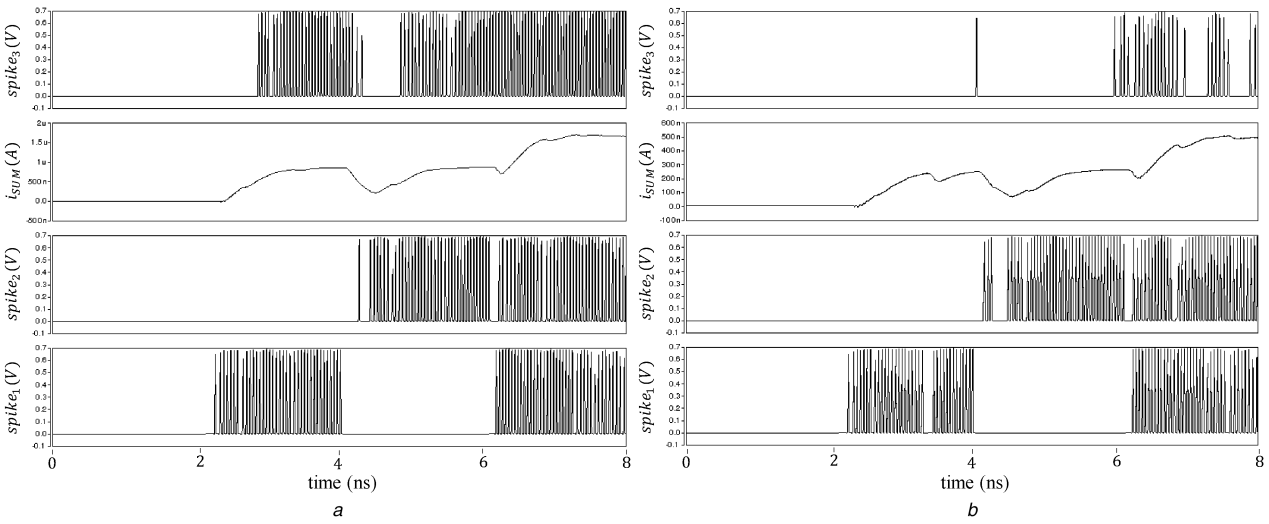

Fig. 7 *Simulation transients of S3N with neuromorphic synapse implementing perceptron functionality*
(a) Two inputs, $spike_1$ and $spike_2$, generate the synaptic current i_{SUM} with a weight of 0.14 V that is used as the input current to the output S3N, whose resulting spike train is $spike_3$, which realises OR logic, ***(b)*** using the same circuit, the weight is adjusted to 0.2 V, and the resulting AND function is realised

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 spintronic device we proposed in this work that combines an MTJ that has a thermally unstable nanomagnet with an HM exhibiting GSHE has not been experimentally demonstrated yet, even though each individual component has been demonstrated by different authors. Two-terminal MTJs having unstable free layers have been experimentally utilised for TRNG applications [43–46] while GSHE-driven MTJs with stable free layers are also commonly demonstrated [33, 34] for memory applications. More recently, an embedded magnetoresistive random access memory-based implementation of a p-bit was proposed that uses a two-terminal MTJ with an unstable free layer along with an NMOS transistor [47]. Our main results would remain essentially unchanged if such an alternative p-bit replaces the three-terminal device proposed herein, and whether a three-terminal device proves to be more flexible due to the separate control terminal deserves further study.

6.2 Extensions and future work

Theoretical neuroscience has demonstrated that networks of stochastic neurons having firing rates which follow a Poisson distribution can achieve Markov Chain Monte-Carlo 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 [48]. 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. Realisation of hardware-based ANN 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 organisation 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 maximisation [49]. 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, utilising the S3N with an appropriate synaptic architecture, one could realise 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 utilisation 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 utilising binary synaptic weights with probabilistic Hebbian learning rules [50–53]. Hence, it should be possible to implement S3Ns with dense arrays of binary stochastically switching memory devices, such as spin-transfer torque (STT)-MTJs or conductive-bridging random access memory (CBRAM), to realise dense and fast unsupervised learning architectures for future cognitive systems.

7 Conclusion

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 0.8–9.6 μW. The circuit, when combined with a neuromorphic second-order synapse, is capable of realising 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 realisation of hardware-based online learning architectures utilising stochastic learning algorithms.

8 Acknowledgments

This work was supported in part by the Center for Probabilistic Spin Logic for Low-Energy Boolean and Non-Boolean Computing (CAPSL), one of the Nanoelectronic Computing Research (nCORE) Centers as task 2759.006, a Semiconductor Research Corporation (SRC) program sponsored by the NSF through CCF 1739635.

9 References

[1] Rosenblatt, F.: 'The perceptron, a perceiving and recognizing automaton Project Para', Cornell Aeronautical Laboratory, 1957

[2] Hodgkin, A. L., Huxley, A.F.: 'A quantitative description of membrane current and its application to conduction and excitation in nerve', *J. Physiol.*, 1952, **117**, (4), pp. 500–544

[3] Koch, C., Segev, I., eds. '*Methods in neuronal modeling: from ions to networks*' (MIT Press, Cambridge, Massachusetts, 1998)

[4] Devadoss, A.V., Ligori, T.A.A.: 'Stock prediction using artificial neural networks', *Int. J. Data Min. Tech. Appl.*, 2013, **2**, pp. 283–291

[5] Kar, A.: 'Stock prediction using artificial neural networks'. Department of Computer Science and Engineering, IIT Kanpur, 1990

[6] Churchland, P.S., Sejnowski, T.J.: '*The computational brain*' (MIT press, Cambridge, Massachusetts, 2016)

[7] Merolla, P.A., Arthur, J.V., Alvarez-Icaza, R., *et al.*: 'A million spiking-neuron integrated circuit with a scalable communication network and interface', *Science*, 2014, **345**, (6197), pp. 668–673

[8] Surace, S.C., Pfister, J.P.: 'A statistical model for in vivo neuronal dynamics', *PLoS One*, 2015, **10**, (11), p. e0142435

[9] Ma, W.J., Beck, J.M., Latham, P.E., *et al.*: 'Bayesian inference with probabilistic population codes', *Nat. Neurosci.*, 2006, **9**, (11), pp. 1432–1438

[10] Deco, G., Rolls, E.T., Romo, R.: 'Stochastic dynamics as a principle of brain function', *Progress Neurobiol.*, 2009, **88**, (1), pp. 1–16

[11] McDonnell, M.D., Ward, L.M.: 'The benefits of noise in neural systems: bridging theory and experiment', *Nat. Rev. Neurosci.*, 2011, **12**, (7), pp. 415–426

[12] Camsari, K.Y., Faria, R., Sutton, B.M., *et al.*: 'Stochastic p-bits for invertible logic', *Phys. Rev. X* **7**, 2017, (3), p. 031014

[13] Yu, T., Cauwenberghs, G.: 'Analog VLSI biophysical neurons and synapses with programmable membrane channel kinetics', *IEEE Trans. Biomed. Circuits Syst.*, 2010, **4**, (3), pp. 139–148

[14] Wijekoon, J.H.B., Dudek, P.: 'Compact silicon neuron circuit with spiking and bursting behaviour', *Neural Netw.*, 2008, **21**, (2), pp. 524–534

[15] Srivastava, S., Rathod, S.S.: 'Silicon neuron-analog CMOS VLSI implementation and analysis at 180nm'. Devices, Circuits and Systems (ICDCS), 2016 3rd Int. Conf., 2016, pp. 28–32

[16] Arthur, J.V., Boahen, K.A.: 'Silicon-neuron design: a dynamical systems approach', *IEEE Trans. Circuits Syst. I Regul. Pap.*, 2011, **58**, (5), pp. 1034–1043

[17] Indiveri, G., Linares-Barranco, B., Hamilton, T.J., *et al.*: 'Neuromorphic silicon neuron circuits', *Front. Neurosci.*, 2011, **5**, pp. 1–23

[18] Grassia, F., Levi, T., Kohno, T., *et al.*: 'Silicon neuron: digital hardware implementation of the quartic model', *J. Artif. Life Robot.*, 2014, **19**, (3), pp. 215–219

[19] Benjamin, B.V., Gao, P., McQuinn, E., *et al.*: 'Neurogrid: a mixed-analog-digital multichip system for large-scale neural simulations', *Proc. IEEE*, 2014, **102**, (5), pp. 699–716

[20] Diehl, P.U., Pedroni, B.U., Cassidy, A., *et al.*: 'Truehappiness: neuromorphic emotion recognition on truenorth'. Neural Networks (IJCNN), 2016 Int. Joint Conf., 2016, pp. 4278–4285

[21] Babacan, Y., Kaçar, F., Gürkan, K.: 'A spiking and bursting neuron circuit based on memristor', *Neurocomputing*, 2016, **203**, pp. 86–91

[22] Fan, D., Sharad, M., Sengupta, A., *et al.*: 'Hierarchical temporal memory based on spin-neurons and resistive memory for energy-efficient brain-inspired computing', *IEEE Trans. Neural Netw. Learn. Syst.*, 2016, **27**, (9), pp. 1907–1919

[23] Sharad, M., Fan, D., Roy, K.: 'Spin-neurons: a possible path to energy-efficient neuromorphic computers', *J. Appl. Phys.*, 2013, **114**, (23), p. 234906

[24] Sharad, M., Fan, D., Roy, K.: 'Ultra low power associative computing with spin neurons and resistive crossbar memory'. Design Automation Conf. (DAC), 2013 50th ACM/EDAC/IEEE, 2013, pp. 1–6

[25] Hu, M., Wang, Y., Wen, W., *et al.*: 'Leveraging stochastic memristor devices in neuromorphic hardware systems', *IEEE J. Emerg. Sel. Topics Circuits Syst.*, 2016, **6**, (2), pp. 235–246

[26] Al-Shedivat, M., Naous, R., Cauwenberghs, G., *et al.*: 'Memristors empower spiking neurons with stochasticity', *IEEE J. Emerg. Sel. Topics Circuits Syst.*, 2015, **5**, (2), pp. 242–253

[27] Wijesinghe, P., Ankit, A., Sengupta, A., *et al.*: 'An all-memristor deep spiking neural network: a step towards realizing the low power, stochastic brain', arXiv preprint arXiv:1712.01472, 2017

[28] Tuma, T., Pantazi, A., Gallo, M.L., *et al.*: 'Stochastic phase-change neurons', *Nat. Nanotechnol.*, 2016, **11**, (8), pp. 693–699

[29] Liyanagedera, C.M., Sengupta, A., Jaiswal, A., *et al.*: 'Magnetic tunnel junction enabled stochastic spiking neural networks: from non-telegraphic to telegraphic switching regime', arXiv preprint arXiv:1709.09247, 2017

[30] Srinivasan, G., Sengupta, A., Roy, K.: 'Magnetic tunnel junction enabled all-spin stochastic spiking neural network'. 2017 Design, Automation & Test in Europe Conf. & Exhibition (DATE), 2017, pp. 530–535

[31] Sengupta, A., Parsa, M., Han, B., *et al.*: 'Probabilistic deep spiking neural systems enabled by magnetic tunnel junction'. *IEEE Trans. Electron Devices*, 2016, **63**, (7), pp. 2963–2970

[32] Sutton, B., Camsari, K.Y., Behin-Aein, B., *et al.*: 'Intrinsic optimization using stochastic nanomagnets', Scientific Reports, 2017, **7**

[33] Liu, L., Pai, C.F., Li, Y., *et al.*: 'Spin-torque switching with the giant spin hall effect of tantalum', *Science*, 2012, **336**, (6081), pp. 555–558

[34] Gosavi, T., Manipatruni, S., Aradhya, S., *et al.*: 'Experimental demonstration of efficient spin-orbit torque switching of an MTJ with sub-100 ns pulses', *IEEE Trans. Magn.*, 2017, **53**, pp. 1–7

[35] Locatelli, N., Mizrahi, A., Accioly, A., *et al.*: 'Noise-enhanced synchronization of stochastic magnetic oscillators', *Phys. Rev. Appl.*, 2014, **2**, (3), p. 034009

[36] Debashis, P., Faria, R., Camsari, K.Y., *et al.*: 'Experimental demonstration of nanomagnet networks as hardware for ising computing'. Electron Devices Meeting (IEDM), 2016 IEEE Int., 2016, pp. 1767–1770

[37] Camsari, K.Y., Ganguly, S., Datta, S.: 'Modular approach to spintronics', Scientific reports, 2015, **5**

[38] Hong, S., Sayed, S., Datta, S.: 'Spin circuit representation for the spin hall effect', *IEEE Trans. Nanotechnol.*, 2016, **15**, (2), pp. 225–236

[39] Zhao, W., Cao, Y.: 'New generation of predictive technology model for sub-45 nm early design exploration', *IEEE Trans. Electron Devices*, 2006, **53**, (11), pp. 2816–2823, http://ptm.asu.edu/

[40] Camsari, K.Y., Ganguly, S., Datta, D., *et al.*: 'Physics-based factorization of magnetic tunnel junctions for modeling and circuit simulation'. Electron Devices Meeting (IEDM), 2014 IEEE Int., 2014, pp. 35–36

[41] Wu, S.M., Cybart, S.A., Yi, D., *et al.*: 'Full electric control of exchange bias', *Phys. Rev. Lett.*, 2013, **110**, (6), p. 067202

[42] Aggarwal, A., Horiuchi, T.K.: 'Neuromorphic VLSI second-order synapse', *Electron. Lett.*, 2015, **51**, (4), pp. 319–321

[43] Choi, W.H., Lv, Y., Kim, J., *et al.*: 'A magnetic tunnel junction based true random number generator with conditional perturb and real-time output probability tracking'. Electron Devices Meeting (IEDM), 2014 IEEE Int., 2014, pp. 12–15

[44] Vodenicarevic, D., Locatelli, N., Mizrahi, A., *et al.*: 'Low-energy truly random number generation with superparamagnetic tunnel junctions for unconventional computing', *Phys. Rev. Appl.*, 2017, **8**, (5), p. 054045

[45] Parks, B., Bapna, M., Igbokwe, J., *et al.*: 'Superparamagnetic perpendicular magnetic tunnel junctions for true random number generators', *AIP Adv.*, 2017, **8**, (5), p. 055903

[46] Lee, H., Ebrahimi, F., Amiri, P.K., *et al.*: 'Design of high-throughput and low-power true random number generator utilizing perpendicularly magnetized voltage-controlled magnetic tunnel junction', *AIP Adv.*, 2017, **7**, (5), p. 055934

[47] Camsari, K.Y., Salahuddin, S., Datta, S.: 'Implementing p-bits with embedded MTJ', *IEEE Electron Device Lett.*, 2017, **38**, (12), pp. 1767–1770

[48] Buesing, L., Bill, J., Nessler, B., *et al.*: 'Neural dynamics as sampling: a model for stochastic computation in recurrent networks of spiking neurons', *PLoS Comput. Biol.*, 2011, **7**, (11), p. e1002211

[49] Nessler, B., Pfeiffer, M., Buesing, L., *et al.*: 'Bayesian computation emerges in generic cortical microcircuits through spike-timing-dependent plasticity', *PLoS Comput. Biol.*, 2013, **9**, (4), p. e1003037

[50] Querlioz, D., Bichler, O., Vincent, A.F., *et al.*: 'Bioinspired programming of memory devices for implementing an inference engine', *Proc. IEEE*, 2015, **103**, (8), pp. 1398–1416

[51] Vincent, A.F., Larroque, J., Locatelli, N., *et al.*: 'Spin-transfer torque magnetic memory as a stochastic memristive synapse for neuromorphic systems', *IEEE Trans. Biomed. Circuits Syst.*, 2015, **9**, (2), pp. 166–174

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

[53] Yu, S., Gao, B., Fang, Z., *et al.*: 'Stochastic learning in oxide binary synaptic device for neuromorphic computing', *Front. Neurosci.*, 2013, **7**, pp. 1–9