An Ultra-Low Power Spintronic Stochastic Spiking Neuron with Self-Adaptive Discrete Sampling

Shadi Sheikhfaal, Steven D. Pyle, Soheil Salehi, and Ronald F. DeMara
Department of Electrical and Computer Engineering, University of Central Florida, Orlando, FL
shadi@knights.ucf.edu, steven.pyle@ucf.edu, soheil.salehi@knights.ucf.edu,
ronald.demara@ucf.edu

***Abstract*— State-of-the-art machine learning models have achieved impressive feats of narrow intelligence, but have yet to realize the computational generality, adaptability, and power efficiency of biological brains. Thus, this work aims to improve current neural network models by leveraging 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. 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, along with behavioral robustness to process variation.**

I. INTRODUCTION

Research towards more brain-like spiking neural networks have typically utilized Leaky-Integrate-and Fire (LIF) models or more complex Hodgkin-Huxley models [1]–[3], which do not intrinsically integrate the ubiquitous stochasticity found in brains [4]–[6]. Although stochasticity can be worked into such models, it comes at the cost of additional circuitry, such as Pseudo-Random-Number Generators (PRNGs) [7]. Some recent works have proposed intrinsically stochastic spiking neuron circuits utilizing emerging devices, such as memristors [8]–[10], phase-change devices [11], or spintronic devices [12]–[15]. However, these circuits typically utilize the stochasticity of the switching behavior for such devices, and thus, requires write-read-reset cycling with extraneous power and delay overheads. Accordingly, the *Subthreshold Spintronic Stochastic Spiking Neuron (S4N)* delineated herein is designed to naturally generate stochastic spiking signals in-situ, at ultra-low-power and high speed.

The organization of the remainder of this paper is as follows. Section II introduces background information relevant to the design of the S4N. Section III describes the S4N circuitry and operation, Section IV contains the results and analysis of the S4N. Section V concludes the paper.

II. BACKGROUND

Neural Sampling is a concept from computational neuroscience that gives weight to the computational abilities of stochastic neurons. Probabilistic spintronics utilizes the thermally-driven stochastic behavior of low-energy-barrier magnetic devices for in-circuit true randomness at high speed. A brief discussion on previous stochastic spiking circuits is provided in this section.

A. Neural Sampling

Neural Sampling is a theoretical framework in computational neuroscience which postulates that the stochastic firing behavior of in-vivo cortical neurons corresponds to samples of an underlying conditional distribution [16]. By leveraging networks of such stochastically sampling neurons, probabilistic inference can be carried out on variables of interest. Furthermore, additional work has demonstrated that stochastic spiking neurons within cortical network motifs combined with Hebbian learning approximates an online version of Expectation Maximization, an effective statistical tool for realizing generative models, which is key for unsupervised learning [17], [18]. Therefore, stochastically spiking neurons implementing Neural Sampling can be a powerful model for achieving unsupervised learning in neuromorphic circuits and architectures, which is not compatible with LIF neurons. Thus, this work aims to enable such neural motifs at ultra-low-power.

B. Probabilistic Spintronics

The spintronic device utilized herein originates from a novel probabilistic adaptation of the Magnetic Tunnel Junction (MTJ) first proposed as part of a 1-Transistor-with-1-MTJ structure called an embedded probabilistic bit (p-bit) [19]. The MTJ of the p-bit stochastically switches between its Anti-Parallel (AP) and Parallel (P) states due to the very low energy barrier (Λ) of the free layer, where the mean retention time for an MTJ (τ) is given by (1).

$$\tau = \tau_0 exp(\Lambda/kT) \tag{1}$$

Where τ_0 is a material dependent parameter called the attempt time, k is Boltzmann's constant, and T is the temperature in Kelvin [19]. The stochastic MTJ (sMTJ) used herein is not the same as the p-bit device. The p-bit device contains specific circuitry besides the sMTJ which is different to what is used for the S4N.

C. Stochastic Spiking Neurons in Hardware

Stochastically spiking neural circuits have been realized using digital CMOS approaches as well as emerging devices of which we review a recent selection. Digital CMOS approaches, such as IBM's TrueNorth chip [7], rely on PRNG circuits for generating stochasticity, which have a large area and energy cost, in addition to lacking true randomness. The work developed in [15] leverages the tunably-stochastic behavior of p-bits to realize a high-speed asynchronous stochastically spiking neuron, but the power is still rather large, and the requirement for nearly-zero energy barrier

978-1-7281-2788-0/19/$31.00 ©2019 IEEE

MTJs is quite strict. Thus, the S4N is introduced herein, which is capable of high-speed stochastic spiking behavior at extremely low power.

III. 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, such that it can be relevant for implementing Neural Sampling. This has lead to a circuit that appears rather different than traditional rate-based spiking neuron schemes where the output is purely a Poissonian spike rate yet, the S4N is still relevant for cortically inspired computations 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 [20], but also assists in balancing the network to be reasonably sensitive, even in the presence of process variation.
4) Process variation effects do not 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 [21].

The S4N circuit shown in Figure 1 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 it is 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 1, 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

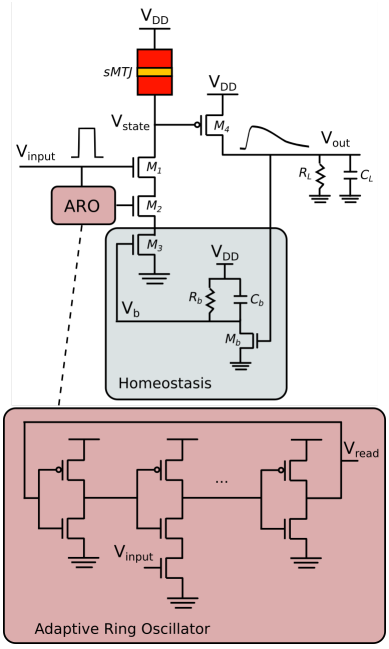

Fig. 1. The S4N circuit developed herein.

operation, an ultra-low-power scheme is realized.

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 the ratio of $(R_{high} - R_{low})/R_{low}$ is typically 100-150% in MTJ devices, 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 1 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.

IV. Results and Analysis

This Section analyzes the results of our simulations, which were performed using HSPICE with high-performance 7nm FinFET PTM Transistor models [22]. The sMTJ was modeled using physically benchmarked spintronic modules from the Modular Spintronic Library [23], [24]. The other circuit parameters are listed in Table I.

TABLE I
CIRCUIT PARAMETERS

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

A. Circuit Analysis

Figure 2 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.

B. Variation Analysis

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

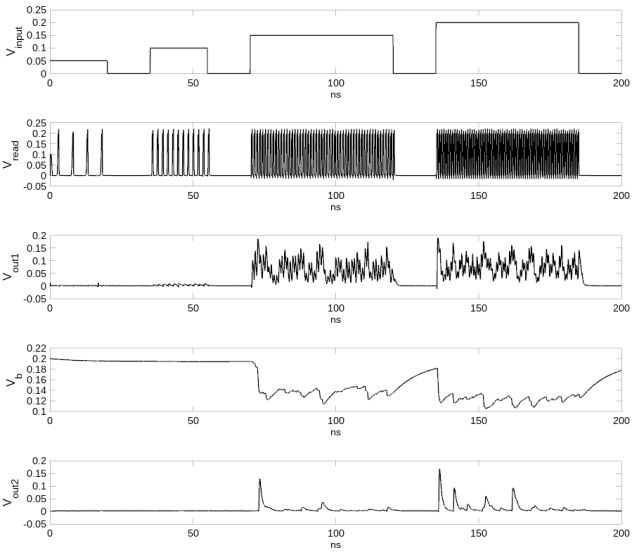

Fig. 2. The operational waveforms of the S4N circuit.

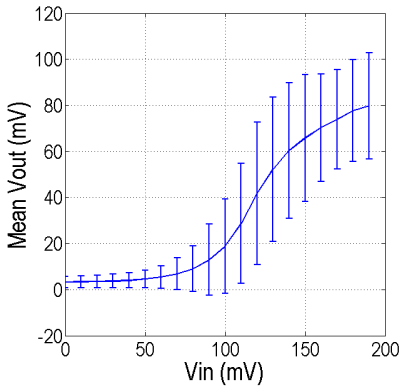

Fig. 3. The mean output voltage of the S4N versus input voltage under process variation.

75mV and all resistances and capacitances listed in Table I with a standard deviation of 20%. As shown in Figure 3, 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. 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 [25].

C. Power Analysis

The S4N circuit, operating at 200mV, in the presence of process variation, uses a maximum power of just 77nW, as shown in Figure 4, which is incredibly efficient for a spiking neuron design operating at the nanosecond timescale. Additionally, the power consumption scales in an almost sigmoidal fashion to the input voltage, using up to about an 8× reduction in power at low input voltages, which would be the most likely operating region for most S4Ns in a large network architecture.

D. Area Overhead

We compared the area overhead on the proposed neuron with different CMOS spiking neurons including the

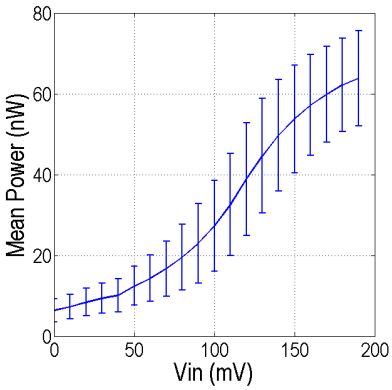

Fig. 4. The mean power consumption of the S4N versus input voltage under process variation.

oscillatory models (Resonate-and-Fire type with 24-T and Hindmarsh-Rose type with more than 40-T) [25], [26], and conductance-based neuron model (Hodgkin-Huxley type with more than 40-T) [27]. The S4N requires only 16 transistors and one sMTJ, which is more compact than the aforementioned neurons. Since the bulk of the device count is due to the ARO, further area reductions could be achieved with novel sampling circuitry used to replace the ARO unit, which is left for future work.

V. CONCLUSIONS

This work 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. Additionally, the presented neuron exhibits area improvements compared to CMOS based neurons. The area and power efficiency of this design is especially important since the neuron circuit is intended to be used in a large-scale VLSI neural networks consisting many thousands of neurons. Such circuit could pave the way to realizing improved efficiency in neuromorphic circuits.

ACKNOWLEDGMENT

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.

REFERENCES

[1] 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.

[2] C. Lee, G. Srinivasan, P. Panda, and K. Roy, "Deep spiking convolutional neural network trained with unsupervised spike timing dependent plasticity," *IEEE Transactions on Cognitive and Developmental Systems*, 2018.

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

[4] W. Maass, "To spike or not to spike: that is the question," *Proceedings of the IEEE*, vol. 103, no. 12, pp. 2219–2224, 2015.

[5] 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.

[6] A. A. Faisal *et al.*, "Noise in the nervous system," *Nature reviews neuroscience*, vol. 9, no. 4, p. 292, 2008.

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

[8] M. Hu, Y. Wang, W. Wen, Y. Wang, and H. Li, "Leveraging stochastic memristor devices in neuromorphic hardware systems," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 6, no. 2, pp. 235–246, 2016.

[9] M. Al-Shedivat *et al.*, "Memristors empower spiking neurons with stochasticity," *IEEE journal on Emerging and selected topics in circuits and systems*, vol. 5, no. 2, pp. 242–253, 2015.

[10] P. Wijesinghe *et al.*, "An all-memristor deep spiking neural computing system: A step towards realizing the low power, stochastic brain," *arXiv preprint arXiv:1712.01472*, 2017.

[11] T. Tuma *et al.*, "Stochastic phase-change neurons," *Nature nanotechnology*, vol. 11, no. 8, p. 693, 2016.

[12] C. M. Liyanagedera *et al.*, "Stochastic spiking neural networks enabled by magnetic tunnel junctions: From nontelegraphic to telegraphic switching regimes," *Physical Review Applied*, no. 8, 064017, 2017.

[13] G. Srinivasan, A. Sengupta, and K. Roy, "Magnetic tunnel junction enabled all-spin stochastic spiking neural network," *DATE*, pp. 530–535, IEEE 2017.

[14] A. Sengupta, M. Parsa, B. Han, and K. Roy, "Probabilistic deep spiking neural systems enabled by magnetic tunnel junction," *IEEE Transactions on Electron Devices*, vol. 63, no. 7, pp. 2963–2970, 2016.

[15] 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 Digital Techniques*, vol. 12, no. 4, pp. 122–129, 2018.

[16] 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, 2011.

[17] 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, 2013.

[18] 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:1707.05182*, 2017.

[19] 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.

[20] 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, pp. 1285–1310, 2001.

[21] D. A. McCormick *et al.*, "Comparative electrophysiology of pyramidal and sparsely spiny stellate neurons of the neocortex," *Journal of neurophysiology*, vol. 54, no. 4, pp. 782–806, 1985.

[22] S. Sinha *et al.*, "Exploring sub-20nm finfet design with predictive technology models," *49th DAC*, pp. 283–288, 2012.

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

[24] R. Zand, A. Roohi, and R. F. DeMara, "Energy-efficient and process-variation-resilient write circuit schemes for spin hall effect mram device," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 25, no. 9, pp. 2394–2401, 2017.

[25] K. Nakada, T. Asai, and H. Hayashi, "A silicon resonate-and-fire neuron based on the volterra system," *Proceeding of the NOLTA*, pp. 82–85, 2005.

[26] Y. J. Lee *et al.*, "Low power real time electronic neuron vlsi design using subthreshold technique," *IEEE ISCAS*, pp. IV–744, 2004.

[27] M. F. Simoni and S. P. DeWeerth, "Adaptation in a vlsi model of a neuron," *IEEE TCAS II: Analog and digital signal processing*, vol. 46, pp. 967–970, 1999.