CMOS-RRAM Neuromorphic Accelerators Using Multi-bit Neurons

Vishal Saxena, and Aly Moussa
Electrical and Computer Engineering Department
University of Delaware
Newark, DE, USA
vsaxena@udel.edu

***Abstract*—Recent foundry-based integration of Resistive Random Access Memory (RRAM) with standard CMOS drives interest in developing high-density, low-power Edge-AI accelerators. 1TnR RRAM arrays have shown promise for realizing very low-energy Vector Matrix Multiplication (VMM) computation in the analog domain. On the other end of the spectrum, spiking neural networks (SNNs) promise low-power computing by eliminating energy-expensive data converters. However, mixed-signal circuit designers must account for RRAM nonidealities and innovate circuits that bridge the performance gap between digital ANNs and analog SNNs. This article reviews this area and presents novel multi-level spiking CMOS neurons that easily interface with RRAMs while providing higher-resolution encoding.**

***Index Terms*—Artificial Intelligence (AI), CMOS Neurons, non-volatile memory (NVM), Neuromorphic computing, Resistive RAM (RRAM).**

I. Introduction

MIXED-signal CMOS circuits built around crossbar analog compute arrays have emerged as a promising enabler for low-power Edge-AI accelerators. A wide range of circuits and devices have been proposed for performing vector-matrix multiplication (**VMM**) in these arrays, leading to convergent paradigms such as compute in memory (**CIM**) and neuromorphic computing. The overarching goal of these architectures is to perform computation inside or close to memory to minimize energy consumed in data transfers between processor and memory, *i.e.* the von Neumann bottleneck.

CIMs using on-chip SRAMs with switched capacitor [1] or continuous-time VMM [2] circuits have demonstrated close to GPU-like inference accuracy while allowing hardware re-use for executing a large artificial neural network (**ANN**) model, such as CIFAR-10/100. Here, the model weights are loaded, and VMM is performed using charge sharing in capacitors, as each ANN layer is partially or fully processed. On the other hand, non-volatile memory arrays (**NVMs**) seek the persistence of model weights with energy consumed only for data movement [3], [4]. Since each model layer needs to be physically manifested in hardware, high-density and ultra-low power operation of these arrays becomes necessary.

While floating gate NVMs have shown proven multi-level weights, Resistive random access memory (**RRAM**), also known as memristors, have been intensely investigated over the past decade [5]–[8]. A foremost challenge with mixed-signal VMMs is the energy and area overhead introduced by digital-to-analog converters (**DACs**) and analog-to-digital converters (**ADCs**) [9]. Secondly, these VMMs incur precision loss due to the PVT-dependent non-linear I-V characteristics of the NVM cells [3], [4], [8], [10].

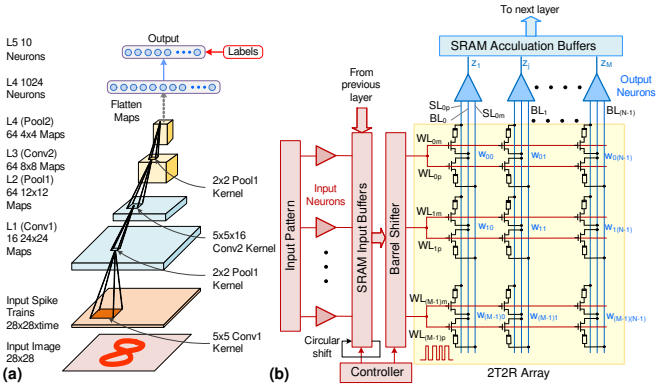

Fig. 1. A spike-based mixed-signal AI accelerator architecture based on signed-1T1R (i.e., 2T2R) RRAM cross-point arrays: (a) A deep SNN with input, Convnet, and output layers of spiking neurons, (b) the 2T2R array for realizing a single fully-connected or CNN layer.

A neural-inspired or ***Neuromorphic computing*** architecture eliminates the need for data converters by using 'spikes' for information encoding, computation, and communication. The resulting spiking neural networks (**SNNs**) employ NVMs as 'synapses' that realize the ANN weights and potentially allow for localized on-chip learning [4], [11]. Recently, asynchronous digital Neuromorphic ICs such as Loihi2 [12] have demonstrated significant gains in energy per inference [13]. These can potentially benefit from mixed-signal realizations should they alleviate the aforementioned challenges.

In this paper, we increase the information capacity of the spiking neurons to represent higher-resolution activation. The rest of the manuscript is organized as follows: Section II briefly describes CMOS-RRAM circuits for spike-domain VMM computation; Section III presents the novel high-resolution spiking CMOS neuron. Finally, Section IV presents simulation results followed by the conclusion.

II. CMOS-RRAM Hardware SNN Design

A. Spike-based Encoding for Linear Synapses

In NVM arrays employed as analog-domain VMMs, each cell must perform four-quadrant analog multiplication with

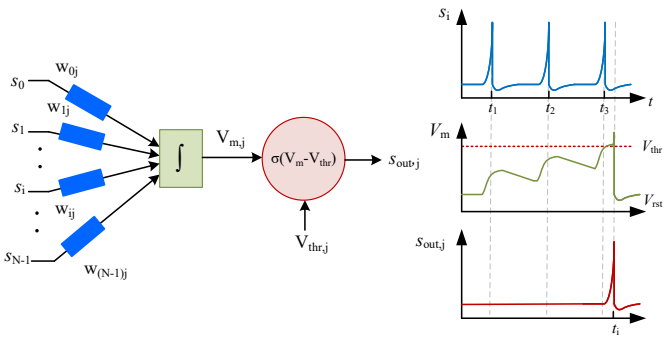

Fig. 2. Illustration of a conventional leaky integrate and fire (LIF) neuron.

sufficient dynamic range. The RRAM cell's I-V nonlinearity distorts this multiplication, thereby reducing the classification accuracy of the ANN [4]. Calibrating the nonlinearity or training around it is challenging due to strong process, voltage, and temperature (**PVT**) dependence.

Pulse or spike-based encoding of ANN activations is employed to mitigate the impact of cell nonlinearity on the accuracy of ANN. Here, an input activation, x, is represented as

$$x \equiv \sum_n V_p x_n p(t - t_n) + \epsilon \tag{1}$$

V_p and $p(t)$ are the spike pulse amplitude and normalized pulse shape. $x_n \in [0, 1]$ is a binary sequence of length with spike times, t_n, and ϵ is the quantization error.

Signed weights (or synapses) are formed by using a 2T2R pair. The select transistor facilitates individual access to the RRAM devices and/or sets the current for forming, Set/Reset, and readout of individual weights.

B. Array Architecture

A mixed-signal ANN accelerator architecture is shown in **Fig. 1** using CMOS-RRAM crossbar arrays. Spike-coded activations not only help linearize the synapses but also eliminate the need for DACs and ADCs. The input spikes to the VMM are either generated using a digital integrate and fire neuron (**IFN**) or buffered spikes from the previous layer. The tensor computation for a CNN layer is mapped to the architecture by time multiplexing the VMM array. While all the feature maps are computed in parallel in the crossbar array [9], [14], array utilization is optimized by implementing the strides by employing rotating input buffers and a barrel shifter; VMM outputs are accumulated in the digital buffers.

C. Spiking Neurons

In the VMM array, the neuron sums and integrates the weighted cell currents on the capacitor, C_m, which results in membrane potential, v_m, that is equivalent to the weighted sum, $y_j = \sum_i w_{ij} x_i$, in a standard ANN. Leakiness of the IFN is not necessary for frame-based computation but can be easily incorporated [15]. The additive noise, η, allows modeling the IFN using a differentiable activation function, $z_j = \sigma(v_{m,j} - V_{thr,j} + \eta)$, to compute *surrogate gradients* for backpropagation based training in SNNs, where $V_{thr,j}^{(l)}$ is the firing threshold and equivalent to the bias term in ANNs [16].

While IFNs replace data converters in an SNN, understanding their effective bit resolution is important. IFNs with a zero refractory period perform lossless spike encoding from which information can be recovered using non-linear methods [17], [18]. However, a finite refractory period introduces non-uniform quantization noise, and a low-pass filter (LPF) can be used for the lossy reconstruction of inputs. IFN parallels asynchronous delta-sigma modulators (ADSM) where an asynchronous comparator with hysteresis is utilized but suffers from limit cycles and doesn't perform noise-shaping [18].

D. SNN Training

Although SNNs can be realized with low-power hardware in the analog domain, the classification accuracy for inference should be comparable with that of digital Edge-AI accelerators. While 4-bit weight resolution is adequate for large models with quantization-aware training, at least 8-bit fixed-point resolution is desired [4]. An early approach was to train the ANN model on GPU using standard Backprop, with techniques including Dropout, and then the weights were normalized and transferred to the equivalent rate-based SNN [19]. Later, surrogate gradients were employed to train SNNs natively using Backprop [16]. In recent work, we demonstrated that ***autograd*** in PyTorch can directly train SNNs with non-differentiable activation and achieve accuracy within 2% of the state-of-the-art on CIFAR10 [14].

III. SNN Accelerator with Multi-bit Neurons

A. Prior Multibit Spiking Neurons

Recently, multi-level neurons were conceptualized to reduce the quantization loss and thus enhance the information content in spike encoding. In a first such effort, M LIF neurons with different threshold voltages were combined to form a single multi-bit neuron unit, and their outputs are summed [20]. This resulted in integer-valued spikes in the range $\{0, 1, \cdots, M\}$.

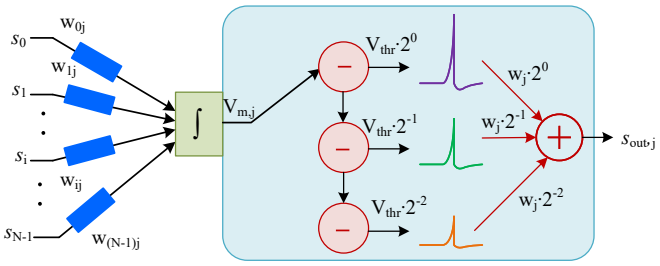

Fig. 3. Conceptual illustration of a multibit spiking neuron [21].

Another recent proposal of multibit neurons is illustrated in **Fig.** 3 where binary (and also fractional) encoding is employed [21]. To analytically understand this, the membrane potential in a LIF neuron is normally distributed in the integration phase. Its probability density function (PDF) is expressed as $f(u) = \frac{1}{\sqrt{(2\pi)}} e^{-u^2/2}$. The associated entropy is then given by

$$H(U) = -\int_{-\infty}^{\infty} f(u) \ln(u) du = \frac{1}{2} \log_2(2\pi e) \tag{2}$$

which is a constant or value 2.047 [21]. When the membrane potential is quantized into spikes, the probabilities of no-spike (0) and spike (1) are given by

$$p_0 = P(u < V_{thr}) = \Phi(V_{thr}) \tag{3}$$
$$p_1 = P(u \geq V_{thr}) = 1 - \Phi(V_{thr}) \tag{4}$$

where Φ is the upper quantile of the normal distribution. Consequently, the entropy of spike encoding is [21]

$$H(S) = -\sum_{i=0,1} p_i \log_2 p_i = -(p_0 \log_2 p_0 + p_1 \log_2 p_1) \tag{5}$$

The information loss of a LIF is

$$L = H(U) - H(S) = \frac{1}{2}\log_2(2\pi e) + p_0 \log_2 p_0 + p_1 \log_2 p_1 \tag{6}$$

which can only be improved by maximizing the spike entropy, $H(S)$ [21].

$$H(S_N) = -\sum_{i=0}^{N-1} p_i \log_2 p_i \tag{7}$$

A careful distribution of firing thresholds can result in 2-bit entropy of 1.36 and 3-bit entropy of 1.998 as opposed to the LIF entropy of 0.848 [21]. This promises a significant reduction in quantization loss in spike encoding.

B. Proposed Multibit Spiking Neuron

We propose multi-bit leaky integrate and fire (**LIF**) neurons to increase the information capacity or the effective number of bits (N_{bit}) in SNN computation. Similar to previous IFNs, this design is asynchronous but with multilevel thresholds. An M-bit asynchronous quantizer with unit-weighted outputs, $D_k \in \{0, 1, \ldots, 2^M - 1\}$ can be realized either in the voltage or time domain. The core idea is to provide a more granular representation of the membrane potential using M integer-weighted spikes with different firing thresholds. The equations can describe the resulting IFN:

$$v_{m,j}(t) = V_{rst} + \sum_{i=1}^{N} w_{ij} \cdot p_i(t) \tag{8}$$

where $p(t) = \int_0^t s(t) \otimes h(t) dt$ is the post-synaptic potential (PSP) due to a single spike, and $h(t) = e^{-t/\tau_m} u(t)$ is the impulse response that incorporates neuron's leaky behavior.

The neuron's binary-coded state, k, depends upon the threshold crossed as $k : V_{ref,k} \leq v_{m,j}(t) < V_{ref,k+1}$. For example, a 2-bit neuron could be in one of the following states:

$$\begin{aligned}
&\text{State 00 (k=0):} \quad V_{bot} \leq v_{m,j}(t) < V_{ref,0} \\
&\text{State 01 (k=1):} \quad V_{ref,0} \leq v_{m,j}(t) < V_{ref,1} \\
&\text{State 10 (k=2):} \quad V_{ref,1} \leq v_{m,j}(t) < V_{ref,2} \\
&\text{State 11 (k=3):} \quad V_{ref,2} \leq v_{m,j}(t) < V_{top}
\end{aligned} \tag{9}$$

After the neuron spikes and a threshold $V_{ref,k}$ is crossed, the membrane potential is reset as per the bit encoding, and the multibit output spike is generated at time t_j^f:

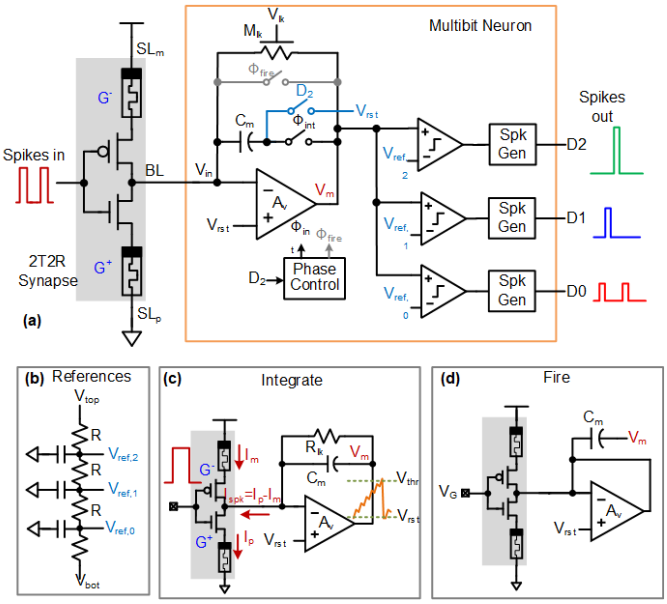

Fig. 4. (a) Schematic of the multi-bit spiking CMOS neuron, (b) reference generation ladder. Opamp reconfiguration in the (c) integration phase and (d) firing phases.

$$v_{m,j}(t) \gtrsim V_{ref,k} : \begin{cases} s_j(t) \leftarrow k \cdot g(t - t_j^f) \\ v_{m,j}(t) \leftarrow V_{rst} \text{ for } k = 2^M - 1 \end{cases} \tag{10}$$

Here, $g(t)$ is the unit spike pulse shape. For higher input spike density, the neuron will likely stay in higher states and fire with higher values spikes. Conversely, without input spikes, the membrane potential decays to $V_{ref,0}$ and the neuron climbs down the ladder of states. Also, the membrane potential is reset to V_{rst} only if the highest index comparator ($k = 2^M - 1$) has an output.

This design also alleviates the issue of diminishing spikes in SNNs, where the number of spikes becomes increasingly sparse as they propagate forward through the layers. As discussed earlier, the 1T1R cells are driven with bilevel pulses, and the thermometer-coded outputs of the $M = 2$-bit IFN are directly used as inputs to the crossbar array. The IFN is reset after each input *frame* of $2^{N_{bit}-M}$ length.

C. IFN Circuit Design

Fig. 4 shows a 2-bit, or 4-level (0 to 3), neuron design. The design re-uses the opamp for inference and array programming and can also be adapted for on-chip learning. This design employs $2^M - 1 = 3$ synchronous comparators and uniform thresholds, V_{thr0-2}, using a resistor ladder with $V_{top} = 1.2V$ and $V_{bot} = 300mV$. The membrane rest potential is $V_{rst} = 100mV$. The asymmetric thresholds allow for the inherent rectified linear unit (**ReLU**) as the non-linear activation function. The integer-coded quantizer outputs, D_0-D_2, can be used directly or combined to form binary-coded output. The integrator is reset only when the spike output, D_2, goes high.

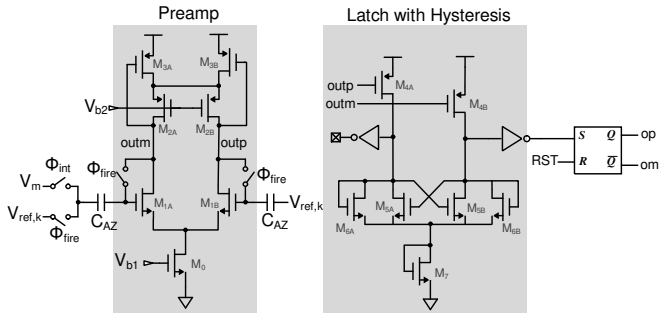

Fig. 5. Asynchronous comparator with auto-zeroing.

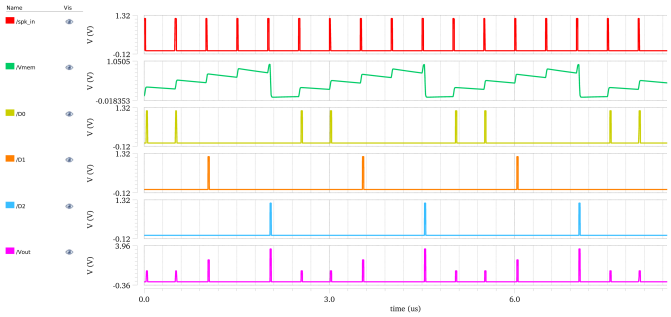

Fig. 6. Transient simulation for the circuit in Fig. 4 demonstrating multi-bit output spikes with 2MHz input spike rate. Here, input spikes are current pulses of 2μA with 20ns pulse width.

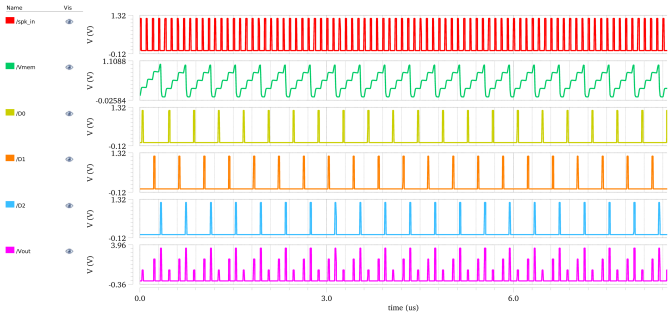

Fig. 7. Transient simulation for multi-bit output spikes with 10MHz input spike rate.

The integrator opamp uses folded-cascode topology with a unity-gain frequency, f_{un}=10MHz, consuming only 6μW of power. Compared with neurons intended for driving resistive RRAM array [15], the G_m-based VMMs require the neuron to only drive the input capacitances of the select transistors, resulting in a simpler digital drive. The asynchronous comparators, shown in **Fig. 5**, is designed using cross-coupled latches with hysteresis, and a pre-amp with offset storage and cancellation. The pre-amp uses a 1μA bias current.

D. Simulation Results

The CMOS neuron shown in **Fig. 4** is designed in TSMC 65nm LP CMOS process with a $V_{DD} = 1.2V$ supply voltage. Cadence Spectre transient simulation result is shown **Fig. 6**, where the output spikes can be seen for a multi-bit neuron with a spiking current input of 2μA at an input spike rate of 2 MHz. We can observe that for lower input pike rate, more D_0 spikes are seen and D_2 is sparse. In contrast in **Fig. 7**, for a 10 MHz input spike rate, all D_k outputs are busy. The analog equivalent output obtained by summing D_k's, $v_{out} = \sum_{k=0}^{3} k \cdot D_k$. We see that v_{out} provides a multi-bit representation of the membrane voltage, with higher accuracy than a single-bit neuron. The performance of the multibit neuron is benchmarked in **Table I**. The energy figure-of-merit is expressed in fJ/spike/synapse for a fan-in and fan-out of 3000 synapses. Thanks to 3-spikes generated per integrator at a maximum output spike rate of 1MHz, the estimated inference energy consumption is 5 fJ/synapse/spike.

While the focus of this work is the multi-bit neuron, preliminary PyTorch-based simulations were performed using the 7-layer convolutional SNN seen in **Fig. 1**. These show a 98.5% accuracy for MNIST and 77.2% accuracy for the CIFAR10 dataset, which improves upon the conventional SNN by 2%.

TABLE I
PERFORMANCE COMPARISON WITH RECENT CMOS NEURONS.

Design	Type	Tech.	Synapse	Bits	I_{bias}	FoM[a]
[15]	Opamp	180nm	1R	1b	13μA	140fJ
[22]	Ring VCO	65nm	None	–	1b	-
[23]	Current sum	65nm	Digital	1b	-	7fJ
[24]	Sub-vT	65nm	None	1b	-	4fJ
[25]	Opamp	180nm	1T1R	1b	9μA	8.1fJ
This work	Opamp	65nm	2T2R	2b	10μA	5fJ

[a] Energy figure-of-merit is expressed as fJ/spike/synapse.

IV. CONCLUSION

While RRAM arrays present a high-density and energy-efficient pathway for designing mixed-signal AI accelerators, the impact of device idealities must be carefully considered. A multi-bit asynchronous spiking neuron increases the quantization resolution or minimizes the information loss in spike encoding while retaining the benefits of spike-based computing. SNNs employing multi-bit neurons exhibit lower accuracy loss than conventional single-bit neurons.

ACKNOWLEDGMENT

This work is supported by the National Science Foundation (NSF) FuSE Award 2329015 and NASA EPSCoR Award 80NSSC22M0171.

REFERENCES

[1] D. Bankman, L. Yang, B. Moons, M. Verhelst, and B. Murmann, "An Always-On 3.8μJ/86% CIFAR-10 Mixed-Signal Binary CNN Processor With All Memory on Chip in 28-nm CMOS," *IEEE Journal of Solid-State Circuits*, vol. 54, no. 1, pp. 158–172, 2018.

[2] J.-O. Seo, M. Seok, and S. Cho, "Archon: A 332.7tops/w 5b variation-tolerant analog cnn processor featuring analog neuronal computation unit and analog memory," in *2022 IEEE International Solid-State Circuits Conference (ISSCC)*, vol. 65, 2022, pp. 258–260.

[3] V. Saxena, "Neuromorphic computing: From devices to integrated circuits," *Journal of Vacuum Science Technology B*, vol. 39, no. 1, p. 010801, 12 2020. [Online]. Available: https://doi.org/10.1116/6.0000591

[4] ——, "Mixed-signal neuromorphic computing circuits using hybrid cmos-rram integration," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 68, no. 2, pp. 581–586, 2021.

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

[6] P. Wang, F. Xu, B. Wang, B. Gao, H. Wu, H. Qian, and S. Yu, "Three-dimensional nand flash for vector–matrix multiplication," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 27, no. 4, pp. 988–991, 2018.

[7] Y. Liao, H. Wu, W. Wan, W. Zhang, B. Gao, H.-S. P. Wong, and H. Qian, "Novel in-memory matrix-matrix multiplication with resistive cross-point arrays," in *2018 IEEE Symposium on VLSI Technology*. IEEE, 2018, pp. 31–32.

[8] M. Prezioso, F. Merrikh-Bayat, B. Chakrabarti, and D. Strukov, "Rram-based hardware implementations of artificial neural networks: progress update and challenges ahead," in *Oxide-based Materials and Devices VII*, vol. 9749. International Society for Optics and Photonics, 2016, p. 974918.

[9] B. Murmann, "Mixed-signal computing for deep neural network inference," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 29, no. 1, pp. 3–13, 2021.

[10] M. R. Mahmoodi and D. Strukov, "An ultra-low energy internally analog, externally digital vector-matrix multiplier based on nor flash memory technology," in *Proceedings of the 55th Annual Design Automation Conference*. ACM, 2018, p. 22.

[11] X. Wu, V. Saxena, and K. Zhu, "Homogeneous spiking neuromorphic system for real-world pattern recognition," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 5, no. 2, pp. 254–266, 2015.

[12] M. Davies, N. Srinivasa, T.-H. Lin, G. Chinya, Y. Cao, S. H. Choday, G. Dimou, P. Joshi, N. Imam, S. Jain, Y. Liao, C.-K. Lin, A. Lines, R. Liu, D. Mathaikutty, S. McCoy, A. Paul, J. Tse, G. Venkataramanan, Y.-H. Weng, A. Wild, Y. Yang, and H. Wang, "Loihi: A neuromorphic manycore processor with on-chip learning," *IEEE Micro*, vol. 38, no. 1, pp. 82–99, 2018.

[13] P. Blouw, X. Choo, E. Hunsberger, and C. Eliasmith, "Benchmarking keyword spotting efficiency on neuromorphic hardware," *arXiv preprint arXiv:1812.01739*, 2018.

[14] A. Dorzhigulov and V. Saxena, "Spiking cmos-nvm mixed-signal neuromorphic convnet with circuit-and training-optimized temporal subsampling," *Frontiers in Neuroscience*, vol. 17, p. 1177592, 2023.

[15] X. Wu, V. Saxena, K. Zhu, and S. Balagopal, "A cmos spiking neuron for brain-inspired neural networks with resistive synapses and in situ learning," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 62, no. 11, pp. 1088–1092, 2015.

[16] E. O. Neftci, H. Mostafa, and F. Zenke, "Surrogate gradient learning in spiking neural networks," *IEEE Signal Processing Magazine*, vol. 36, pp. 61–63, 2019.

[17] N. C. Sevuktekin, L. R. Varshney, P. K. Hanumolu, and A. C. Singer, "Signal processing foundations for time-based signal representations: Neurobiological parallels to engineered systems designed for energy efficiency or hardware simplicity," *IEEE Signal Processing Magazine*, vol. 36, no. 6, pp. 38–50, 2019.

[18] A. S. Alvarado, M. Rastogi, J. G. Harris, and J. C. Principe, "The integrate-and-fire sampler: A special type of asynchronous σ-δ modulator," in *2011 IEEE International Symposium of Circuits and Systems (ISCAS)*. IEEE, 2011, pp. 2031–2034.

[19] P. U. Diehl, D. Neil, J. Binas, M. Cook, S.-C. Liu, and M. Pfeiffer, "Fast-classifying, high-accuracy spiking deep networks through weight and threshold balancing," in *International Joint Conference on Neural Networks (IJCNN)*, 2015, pp. 1–8.

[20] L. Feng, Q. Liu, H. Tang, D. Ma, and G. Pan, "Multi-level firing with spiking ds-resnet: Enabling better and deeper directly-trained spiking neural networks," *arXiv preprint arXiv:2210.06386*, 2022.

[21] Y. Xiao, X. Tian, Y. Ding, P. He, M. Jing, and L. Zuo, "Multi-bit mechanism: A novel information transmission paradigm for spiking neural networks," *arXiv preprint arXiv:2407.05739*, 2024.

[22] B. D. Sahoo, "Ring oscillator based sub-1v leaky integrate-and-fire neuron circuit," in *Circuits and Systems (ISCAS), 2017 IEEE International Symposium on*. IEEE, 2017, pp. 1–4.

[23] B. Larras, P. Chollet, C. Lahuec, F. Seguin, and M. Arzel, "A 65-nm CMOS 7fJ per synaptic event clique-based neural network in scalable architecture," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, 2017, pp. 1–4.

[24] I. Sourikopoulos, S. Hedayat, C. Loyez, F. Danneville, V. Hoel, E. Mercier, and A. Cappy, "A 4-fj/spike artificial neuron in 65 nm cmos technology," *Frontiers in Neuroscience*, vol. 11, p. 123, 2017.

[25] V. Saxena, "A process-variation robust rram-compatible cmos neuron for neuromorphic system-on-a-chip," in *2020 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2020, pp. 1–5.