NeuDW-CIM: a 65-nm 0.8-pJ/Sop Reconfigurable Neuromorphic Compute-in-Memory Macro with Nonlinear Dendrites and K-Winners

Junyi Yang[1], Yahan Yang[1], Shuai Dong[1], Biyan Zhou[1], Ye Ke[1], Zhengnan Fu[1], Xin Si[2], An Guo[1], Peng Zhou[1], Arindam Basu[1*]
[1]Department of Electrical Engineering, City University of Hong Kong, Hong Kong, China
[2]Southeast University, Nanjing, China. (*Corresponding email: arinbasu@cityu.edu.hk)

***Abstract*—This work presents NeuDW-CIM, a highly efficient neuromorphic Compute-in-Memory (CIM) macro for Spiking Neural Networks (SNNs) implemented in 65 nm CMOS. The design introduces a custom twin 9T bit-cell for ternary inputs/weights and a reconfigurable non-linear In-Memory ADC (IMA). The macro supports two specialized modes: 1) Nonlinear Dendrite (NLD) mode, which utilizes reconfigurable IMA to emulate biological dendritic functions, achieving measured accuracies of 97.2% on N-MNIST and 95.5% on DVS Gesture; and 2) Top-K Winner (KWN) mode, featuring an early-stopping mechanism that reduces IMA conversion latency by 30% and digital LIF latency by 10×. Benefiting from the sparse update in KWN mode, NeuDW-CIM achieves a measured energy efficiency (EE) of 0.8 pJ/SOP (1.6× improvement).**

***Index Terms*—Spiking neural networks, SRAM, Computing in-memory, Winner take all, Nonlinear dendrite.**

I. Introduction

Neuromorphic SNNs have proven to be very energy-efficient due to their sparse, binary activations [1], [2] and can provide high-speed, low-latency operation when coupled with dynamic vision sensors (DVS) [3]. Recent work [4] has combined the efficiency of analog CIM for synaptic operations and event-driven clock gating for saving dynamic energy. However, using analog circuits with capacitor based leaky integrate and fire (LIF) neurons reduces reconfigurability. Digital LIF following analog CIM provides flexibility, but faces several other architectural challenges (Fig. 1 (a)). First, it requires an analog-to-digital Converter (ADC) after the analog Multiply-Accumulate (MAC), which increases power and area overheads [5], [6]. Moreover, it suffers from poor throughput due to serial membrane potential (V_{mem}) update of the digital LIF. Third, the output of event sensors [3] is normally ternary with ON/OFF events, which require more channels when mapped to conventional binary input CIM. Lastly, the accuracy of SNN on a limited-precision analog CIM has been limited.

To overcome these issues, a Neuromorphic CIM with Dendrites and Winners (NeuDW-CIM) (Fig. 1 (b)) supporting top-K winner (KWN) and Non-linear (NL) dendrites (NLD) modes is proposed with the following features: **1) NL reconfigurable in-memory ADC (IMA) with small area overhead (12%):** implementing different NL activations with measured average error <1 LSB, achieving high accuracy in NLD mode.

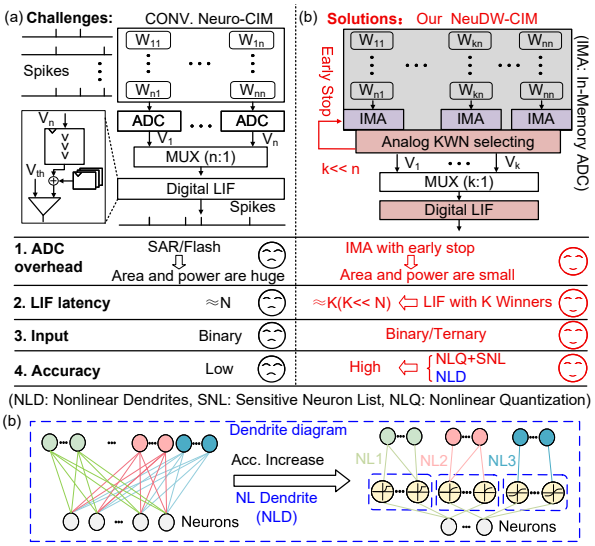

Fig. 1. (a) Challenges and (b) solutions of neuromorphic CIM. (c) Dendrite for increasing accuracy.

2) KWN with early stop: A KWN selection with an early stop mechanism reducing IMA latency by 30% and digital LIF latency by 10×, boosting energy efficiency by 1.6×. **3) Custom twin 9T bit-cell with multi-VDD:** efficiently supporting ternary inputs/weights with enhanced weight precision and minimal overhead. **4) Accuracy optimization:** integrating NL quantization (NLQ) and a sensitive neuron list (SNL) to enhance inference accuracy in KWN mode, with silicon results showing a measured NLQ mean error of 0.41 LSB.

II. Proposed NeuDW-CIM Architecture

Fig. 2 depicts the architecture of the NeuDW-CIM with two modes. It consists of 256×128 and 46×128 static random-access memory (SRAM) arrays for MAC and NL-IMA, respectively. The digital controller configures the ramp ADC for dual task of NL quantization and selection of top-K largest MAC results (K≪128) in KWN mode. Only these indices are used to update V_{mem} for the digital LIF, and the ramp is

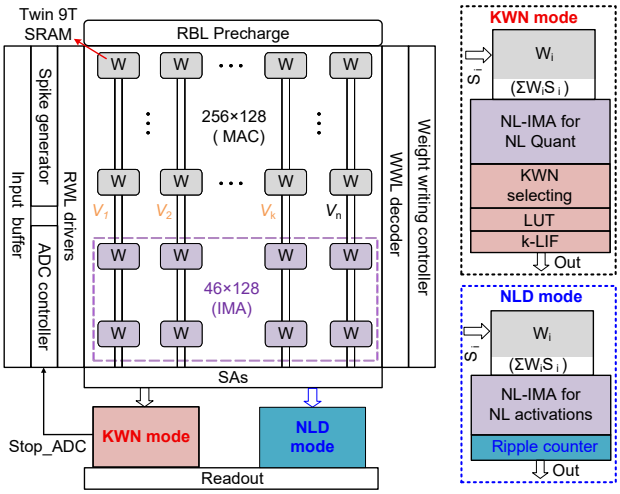

Fig. 2. Overall architecture of the proposed macro and two modes.

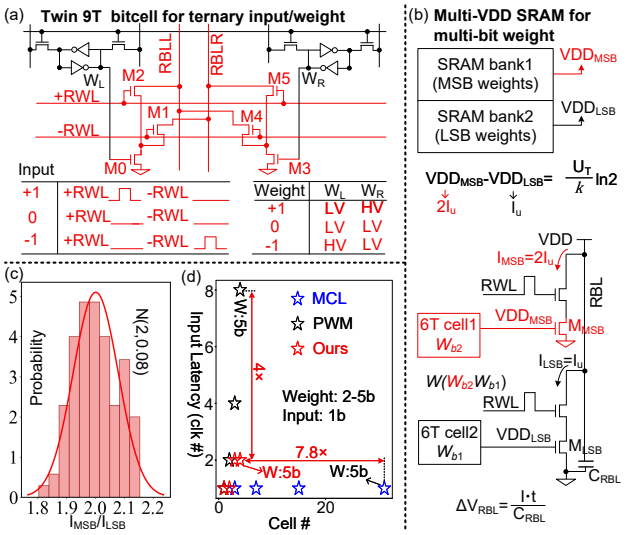

Fig. 3. (a) Custom twin 9T SRAM bit-cell. (b) Multi-VDD SRAM for multi-bit weight. (c) MC simulation of the current ratio. (d) Overhead comparison for implementing multi-bit weight.

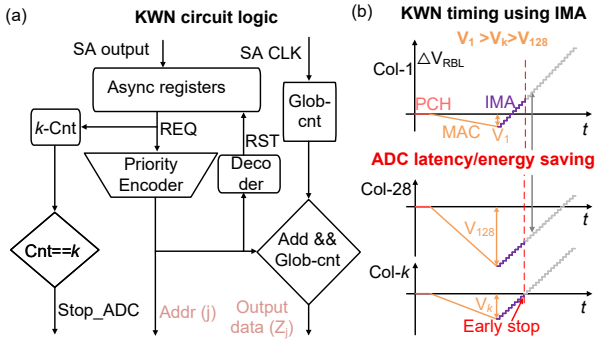

Fig. 4. (a) Circuit controller and (b) timing of the KWN selecting module.

stopped early after the first K conversions resulting in saving ADC/LIF latency and energy. Equation (1) describes the V_{mem} update logic in KWN mode featuring a top-K selection mechanism, where only the K winner neurons undergo membrane potential updates to effectively reduce LIF conversion latency.

$$V^p_{mem}(\mathrm{t}+1) = \begin{cases} (\sum_i W_{i,p}S_i) + \beta V^p_{mem}(\mathrm{t}) + \mathrm{n(t)} \\ \text{(p in top-K)} \\ V^p_{mem}(\mathrm{t}) \quad \text{(p not in top-K)} \end{cases} \quad (1)$$

In NLD mode, various biological dendritic functions with NL activations are integrated into the network to enhance accuracy as illustrated in Fig. 1(c). These NL activations are implemented by the NL-IMA, where outputs from the ramp-IMA are digitized via a ripple counter. Owing to the inherent sparsity of the connections, this enhancement is achieved without increasing the total parameter overhead. Equation (2) defines the update logic of neurons emulating biological NL dendritic characteristics using $f()$, which enhances the learning capability of SNN neurons and improves overall system accuracy.

$$V^p_{mem}(\mathrm{t}+1) = \sum_j W^d_{j,p} f\left(\sum_i W^s_{i,j,p} S_i\right) + \beta V^p_{mem}(\mathrm{t}) \quad (2)$$

A. Twin 9T bit-cell with Multi-VDD Array

Fig. 3(a) shows the schematic of a custom twin 9T SRAM bit-cell that enables multiplication between a ternary input (represented by a pair of +/- RWL signals) and a ternary weight (represented by two 6T-SRAM bit-cells). The SRAM array is partitioned into two banks, each powered by a dedicated supply voltage: VDD_{MSB} and VDD_{LSB} as shown in Fig. 3(b), and the two VDDs are implemented using external power inputs, incurring a minimal power overhead of only 3.5 μW. This Multi-VDD scheme maps the stored weight's MSB and LSB to two distinct discharge currents, I_{MSB} and I_{LSB}, with a fixed ratio such as $I_{MSB} = 2I_{LSB}$ for 3-bit weights. Monte Carlo (MC) simulations show minimal fluctuation in the current ratio (I_{MSB}/I_{LSB}) in Fig. 3(c). We also compare the latency and bit-cell count advantage of our method with conventional pulse-width modulation (PWM) [7], and multi-cell (MCL) [8] approaches for implementing multi-bit weight in Fig. 3(d), showing 4× and 7.8× advantages for the 5-bit weight case.

B. KWN/NLD modes

Fig. 4 depicts the KWN circuit controller and IMA timing, where the IMA is used to create a differential ramp voltage on the read bitlines (RBLs) by sequentially turning on rows after an initial result of MAC is stored on the RBL. Once zero-crossings are detected on the first K RBLs, the controller stops the ADC array according to the early stopping mechanism ($Stop\_ADC$ in Fig. 4(a)). The early stopping mechanism achieves a 30% reduction in ADC conversion latency for the DVS Gesture dataset. For each crossing, the column index, j,

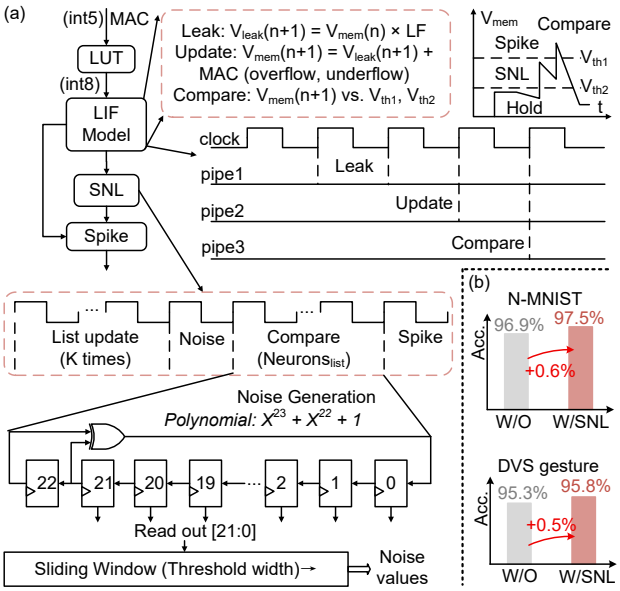

Fig. 5. (a) Flow chart and timing of the LIF module with noise. (b) Accuracy improvement with SNL.

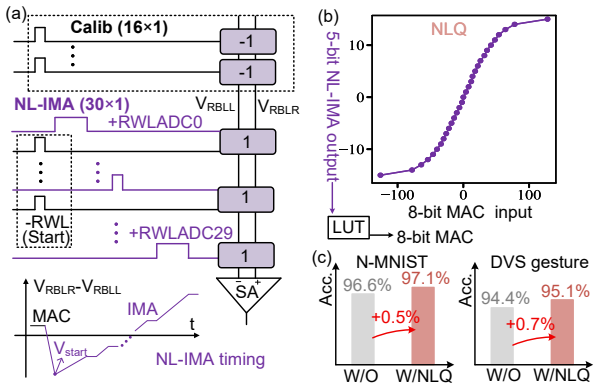

Fig. 6. (a) NL-IMA timing with calibration. (b) NL-IMA for NLQ. (c) Accuracy improvement with NLQ.

obtained from a priority encoder (PENC) and counter value, Z_j (corresponding to quantized MAC), are stored in registers as outputs.

Fig. 5(a) shows the update of the digital LIF in the KWN mode comprising 3 pipelined processes of leak, update and compare. Since only K non-zero Z_j are received from the crossbar, it is possible to introduce large errors in spike times of neurons close to the threshold V_{th1}. Hence, a SNL is maintained such that the V_{mem} of these neurons satisfy $V_{th2} < V_{mem} < V_{th1}$. Further, a Pseudo-Random Binary Sequence (PRBS) is used to generate the noise term n(t) in (1) which enables neurons in SNL to probabilistically fire a spike. SNL and noise addition result in 0.5-0.6% accuracy improvements in two datasets in Fig. 5(b).

The ramp IMA can be operated in NL mode by using

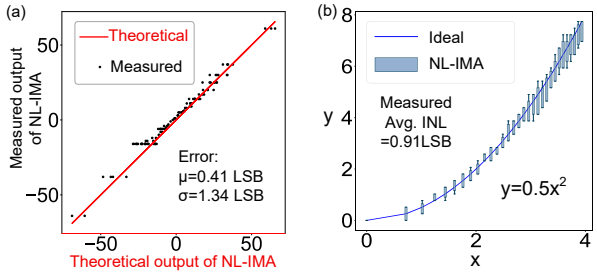

Fig. 7. Measurement results of NL-IMA for (a) NLQ and (b) NL activation.

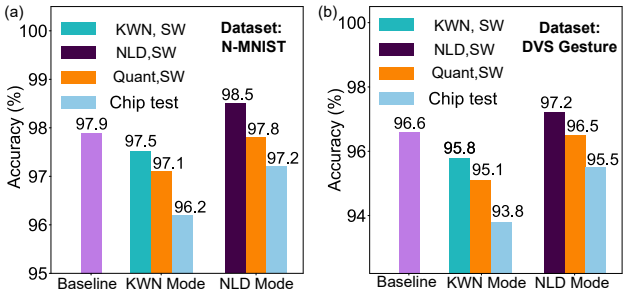

Fig. 8. Accuracy under KWN, NLD modes using 3-bit weight and 5-bit NL-IMA. (a) N-MNIST dataset. (b) DVS Gesture dataset.

variable pulse width for each row (Fig. 6(a), blue +RWLADC). This can be used to create an NL quantization of the MAC as shown in Fig. 6(b) in KWN mode, which helps to reduce ADC resolution (e.g., 5-bit ADC for 8-bit range) and thus energy/latency as well. These NL values are helpful for increasing accuracy (0.5-0.7%) for two datasets in Fig. 6(c) when used in training. In KWN mode, they are mapped back to 8-bit values using a Look-up table (LUT).

In NLD mode, the reconfigurable NL-IMA approximates various NL activations ($f()$ in (2)) by modulating the pulse width of each quantization step as shown in Fig. 6(a), effectively enhancing neuronal learning capabilities across diverse neuromorphic tasks.

III. Measurement Results

NeuDW-CIM is fabricated in 65 nm CMOS, and the die micrograph is shown in Fig. 9(c), which occupies 920 μm×1200 μm with an active area of 0.513 mm^2 (570 μm×900 μm).

Fig. 7(a) compares the measured outputs of NL-IMA for NLQ against theoretical values. The measured data points exhibit a strong correlation with the ideal transfer curve, showing a minor mean error (μ) of 0.41 LSB and a standard deviation (σ) of 1.34 LSB. Furthermore, Fig. 7(b) demonstrates the reconfigurability of NL-IMA in NLD mode by implementing a quadratic activation function ($y = 0.5x^2$). The NL-IMA output closely tracks the ideal value with a measured average integral non-linearity (INL) of only 0.91 LSB.

Fig. 8 shows accuracies on the N-MNIST and DVS Gesture datasets with 3-bit weight and 5-bit NL-IMA. The NLD mode achieves high measured accuracies of 97.2% and 95.5% on

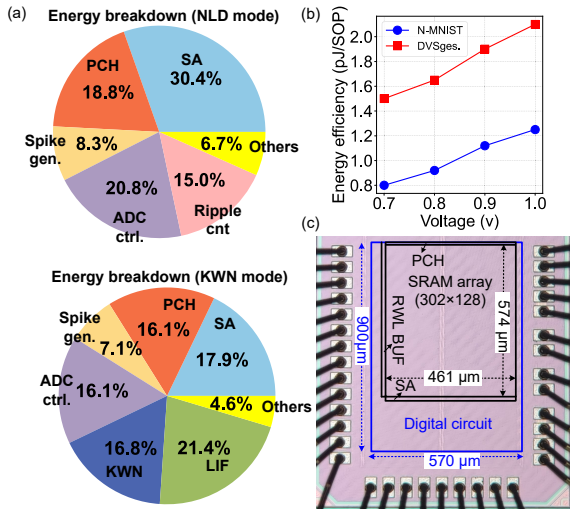

Fig. 9. (a) Energy consumption breakdown for NLD and KWN modes. (b) Energy efficiency across various supply voltages. (c) Die micrograph.

TABLE I
COMPARISON WITH THE STATE-OF-THE-ART NEUROMORPHIC CIM

	ESSERC'25 [2]	ISSCC'23 [1]	ISSCC'24 [4]	VLSI'25 [9]	This Work KWN	This Work NLD
Technology(nm)	65	28	22	130	**65**	
Core area(mm²)	0.99	1.27	2.3	6.75	**0.513**	
Main supply(V)	0.9-1.2	0.56-0.9	0.55-0.9	0.9-1.2	**0.7-1.0**	
Frequency(MHz)	50-150	40-210	51-280	2.5	**50-100**	
Scheme	Digital CIM	Digital	Mixed CIM	Analog CIM	**Analog CIM**	
Macro size	128×48	-	64×128	32×256	**256×128**	
LIF	Digital	Digital	Analog	Analog	**Digital**	
Input	Binary	Binary	Binary	Binary	**Binary/Ternary**	
Weight bit	4/6/8	8	4/8	4	**2-3**	
V_{mem} bit	7/11/15	-	-	-	**12**	
Dataset	- DVS Ges. -	N-MNIST DVS Ges. SeNic	N-MNIST DVS Ges. -	N-MNIST DVS Ges. -	**N-MNIST DVS Ges. Quiroga [10]**	
Accuracy	- 93.54% -	96% 92% 95.7%	97% 94% -	97.1% 90.12% -	**96.2% 93.8% -**	**97.2% 95.5% 96.1%**
Power (mW)	4.9	2.91	0.52	0.096	**0.22**	**0.17**
Energy Efficiency[(a)] (pJ/SOP)	-	1.5 @0.56V	3.78 @0.55V	1.3 @0.9V	**0.8/1.5/- @0.7V**	**1.8/2.3/2.1 @0.7V**

(a) In KWN mode, 0.8 pJ/SOP is for N-MNIST (K=3) and 1.5 pJ/SOP is for DVS Gesture (K=12). In NLD mode, 1.8/2.3/2.1 pJ/SOP corresponds to three datasets with different activations and spike rates.

two benchmark datasets, respectively. Meanwhile, the KWN mode also attains competitive measured results (96.2% and 93.8%, respectively) by incorporating the proposed SNL and NLQ methods.

Fig. 9(a) depicts the energy breakdown of the two modes at VDD = 0.7 V. Measured EEs of K=3 (N-MNIST) and K=12 (DVS Gesture) under different VDD are shown in Fig. 9(b). At VDD = 0.7 V and K=3, NeuDW-CIM achieves the best EE of 0.8 pJ/SOP in KWN mode. The proposed reconfigurable IMA supports both linear and NL modes with a minimal area overhead of only 12%.

Table I compares NeuDW-CIM with other recent works [1], [2], [4], [9]. This work achieves better accuracies and EEs on multiple datasets (97.2% on N-MNIST at 1.8 pJ/SOP, 95.5% on DVS Gesture at 2.3 pJ/SOP) in NLD mode. The early stopping feature of the KWN mode enables higher EEs of 0.8 pJ/SOP (1.6× improvement over the SOTA result 1.3 pJ/SOP in [9]) for K=3 (N-MNIST) and 1.5 pJ/SOP for K=12 (DVS Gesture). While a traditional digital LIF requires serial updates for the entire 128 neurons, our KWN-based sparse update reduces the LIF latency by 10×, as only the K=12 identified winners necessitate V_{mem} update. Furthermore, the additional KWN control logic for early stopping accounts for merely 16.8% of the total power as shown in Fig. 9(a). To show versatility, we tested the macro on a ternary input spike detection dataset (Quiroga [10]) and achieved 2.1 pJ/SOP EE, 96.1% accuracy in NLD mode.

IV. Conclusion

In this work, NeuDW-CIM offers a power-efficient solution for neuromorphic applications via a reconfigurable NL-IMA supporting NLD and KWN modes. Utilizing a multi-V_{DD} twin 9T bit-cell for ternary operations and an early-stopping mechanism, the macro minimizes area overhead for implementing multi-bit weight while reducing ADC and digital LIF latencies by 30% and 10×. NLQ and SNL are utilized to improve inference accuracy in KWN mode. Ultimately, the 65-nm prototype achieves high energy efficiency and inference accuracy on three datasets.

References

[1] J. Zhang *et al.*, "anp-i: A 28nm 1.5 pJ/sop asynchronous spiking neural network processor enabling sub-o. 1 μJ/sample on-chip learning for edge-AI applications," in *2023 IEEE International Solid-State Circuits Conference (ISSCC)*. IEEE, 2023, pp. 21–23.

[2] D. Sharma *et al.*, "A 65nm 5 TOPS/W Digital CIM Accelerator with Reconfigurable Precision and Temporal Pipelining for Spiking Neural Networks," in *2025 IEEE European Solid-State Electronics Research Conference (ESSERC)*. IEEE, 2025, pp. 5–8.

[3] A. Niwa *et al.*, "A 2.97 μm-pitch event-based vision sensor with shared pixel front-end circuitry and low-noise intensity readout mode," in *2023 IEEE International Solid-State Circuits Conference (ISSCC)*. IEEE, 2023, pp. 4–6.

[4] Y. Liu *et al.*, "a 22nm 0.26 nw/synapse spike-driven spiking neural network processing unit using time-step-first dataflow and sparsity-adaptive in-memory computing," in *2024 IEEE International Solid-State Circuits Conference (ISSCC)*, vol. 67. IEEE, 2024, pp. 484–486.

[5] B. Choi *et al.*, "AR-CIM: A 460.22 TOPS/W SRAM-based Analog Reconfigurable Computing-in-Memory Macro with 1/2/4/8-Bit Variable Precision," in *2024 IEEE European Solid-State Electronics Research Conference (ESSERC)*. IEEE, 2024, pp. 361–364.

[6] S. Kim *et al.*, "Neuro-CIM: ADC-less neuromorphic computing-in-memory processor with operation gating/stopping and digital–analog networks," *IEEE Journal of Solid-State Circuits*, vol. 58, no. 10, pp. 2931–2945, 2023.

[7] S. Dong *et al.*, "Topkima-Former: Low-Energy, Low-Latency Inference for Transformers Using Top-k In-Memory ADC," *IEEE Transactions on Circuits and Systems I: Regular Papers*, 2025.

[8] C. Yu *et al.*, "A 65-nm 8T SRAM compute-in-memory macro with column ADCs for processing neural networks," *IEEE Journal of Solid-State Circuits*, vol. 57, no. 11, pp. 3466–3476, 2022.

[9] H. Fu *et al.*, "NeuC-CIM: A 1.3 pJ/SOP Neuromorphic Charge-Domain Compute-in-Memory Macro for Spiking Neural Network," in *2025 Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)*. IEEE, 2025, pp. 1–3.

[10] A. Akhoundi *et al.*, "A Scalable 1024-Channel Ultra-Low-Power Spike Sorting Chip With Event-Driven Detection and Spatial Clustering," *IEEE Journal of Solid-State Circuits*, 2025.