Toward High Performance, Programmable Extreme-Edge Intelligence for Neuromorphic Vision Sensors utilizing Magnetic Domain Wall Motion-based MTJ

Md Abdullah-Al Kaiser[*]
Gourav Datta[*]
University of Southern California
Los Angeles, CA, USA

Peter A. Beerel
University of Southern California
Los Angeles, CA, USA

Akhilesh R. Jaiswal
University of Wisconsin-Madison
Madison, WI, USA

ABSTRACT

The desire to empower resource-limited edge devices with computer vision (CV) must overcome the high energy consumption of collecting and processing vast sensory data. To address the challenge, this work proposes an energy-efficient non-von-Neumann in-pixel processing solution for neuromorphic vision sensors employing emerging (X) magnetic domain wall magnetic tunnel junction (MDWMTJ) for the first time, in conjunction with CMOS-based neuromorphic pixels. Our hybrid CMOS+X approach performs in-situ massively parallel asynchronous analog convolution, exhibiting low power consumption and high accuracy across various CV applications by leveraging the non-volatility and programmability of the MDWMTJ. Moreover, our developed device-circuit-algorithm co-design framework captures device constraints (low tunnel-magnetoresistance, low dynamic range) and circuit constraints (non-linearity, process variation, area consideration) based on monte-carlo simulations and device parameters utilizing GF22nm FD-SOI technology. Our experimental results suggest we can achieve an average of 45.3% reduction in backend-processor energy, maintaining similar front-end energy compared to the state-of-the-art and high accuracy of 79.17% and 95.99% on the DVS-CIFAR10 and IBM DVS128-Gesture datasets, respectively.

KEYWORDS

Neuromorphic, Convolution, In-pixel Processing, Magnetic Domain Wall MTJ, Device-Circuit-Algorithm Co-design.

1 INTRODUCTION

Edge devices with computer vision (CV) systems face energy inefficiency and throughput bottlenecks due to physically segregated sensor hardware and processing platforms [5]. To address this, researchers have explored near-sensor, in-sensor, and in-pixel processing methods [8, 9, 13, 18, 33]. While near-sensor and in-sensor approaches still encounter bandwidth challenges, in-pixel processing allows simultaneous sensing and computing within the pixel array, providing energy-efficient processing by transmitting feature outputs instead of raw sensory data. In the realm of CMOS Image Sensors (CIS), bio-inspired event-based neuromorphic vision sensors (NVS) [19, 22] have gained popularity over traditional CIS for various neural network (NN) applications [6, 24, 27] due to their lower latency, energy efficiency, high temporal precision, and dynamic range. In summary, integrating neuromorphic vision sensors with in-pixel processing presents a comprehensive solution to energy inefficiency and throughput bottlenecks in resource-constrained edge devices with CV systems.

Neuromorphic vision sensors often utilize spiking convolutional neural networks (CNNs) for processing asynchronous input events. In the traditional approach, the duration of the neuromorphic datasets is segmented into predetermined integration intervals, accumulating input spikes within each period to generate multi-bit inputs for the initial layer of the spiking CNN [7]. Unlike subsequent layers that consist of energy-efficient accumulators, the first layer involves digital multi-bit Multiply-Accumulate (MAC) operations. To enhance energy efficiency in the first layer, analog MAC units, employing continuous variables like current, resistance, or pulse width as weights and capacitors as accumulators (representing the membrane potential), can be utilized in CNN hardware implementations [12, 13, 17, 18, 32, 37]. Substantial energy consumption may result from the active amplifier-based capacitor; in contrast, the passive capacitor-based accumulator yields low-energy consumption. However, the passive capacitor cannot retain the charge (membrane potential) for a long duration due to the leakage of the CMOS circuits, resulting in lower overall classification accuracy.

Due to the limitations of the CMOS circuits including the inherent leakage, research is shifting towards NN architectures utilizing emerging (X) post-CMOS technologies such as Phase Change Memories (PCM), Resistive Random Access Memory (RRAM), Magnetic Tunnel Junction (MTJ), Magnetic Domain Wall (MDW), etc. [3, 20, 29, 31]. The benefits of non-volatility, reduced power consumption, higher density, speed, and CMOS compatibility drive this shift. Spintronics devices offer lower latency, reduced energy dissipation, and unlimited endurance compared to other emerging technologies; however, they suffer from a low on-off ratio due to low TMR (between 200% and 600% [1, 15]). Spintronic device magnetic domain wall magnetic tunnel junction (MDWMTJ) demonstrates a continuous resistance state based on domain wall position and has experimentally demonstrated good accuracy in neuromorphic applications [1, 20]. In addition, hybrid CMOS and MDWMTJ structures are reported for logic, in-memory, and neuromorphic applications [14, 31, 34]. Considering these advantages, our proposed

[*] Both authors contributed equally to this research.

DAC '24, June 23–27, 2024, San Francisco, CA, USA
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0601-1/24/06.
https://doi.org/10.1145/3649329.3657359

processing-in-pixel-in-memory (P^2M) hardware for neuromorphic vision sensors utilizes MDWMTJ as the core component. Note, we choose spin-orbit-torque (SOT)-based MDWMTJ instead of spin-transfer-torque (STT)-based MDWMTJ due to their lower write current requirements and decoupled read and write path, resulting in low power dissipation and constant resistance along the write path that makes the associated CMOS circuit design easier.

This work presents an energy-efficient in-pixel processing hardware for spiking CNN focusing on neuromorphic vision applications. We utilize hybrid CMOS and MDWMTJ approaches to achieve high dynamic range, low energy consumption, and programmability for our neuromorphic CMOS+X P^2M hardware. The key contributions of our work include the following:

(1) We propose a novel hybrid CMOS+X approach of processing-in-pixel-in-memory (P^2M) for neuromorphic vision sensors, incorporating the emerging (X) magnetic domain wall magnetic tunnel junction (MDWMTJ) device as the core compute element for our developed spiking CNN framework. The non-volatility and programmability of the MDWMTJ enable retaining the membrane potential for a longer integration time and tunable weight and neuron threshold, which are important for ensuring high accuracy.
(2) In addition, we design three current-based analog weight configurations: CMOS-based, MDWMTJ-based, and hybrid CMOS+X. Our hybrid symbiotic CMOS+X approach combines the unique benefits of both technologies, exhibiting close to state-of-the-art (SOTA) accuracy (thanks to the high dynamic range supported by the CMOS weights) while allowing mapping of multiple CV applications onto the same hardware (owing to the programmability of MDWMTJ).
(3) Finally, we develop a device-circuit-algorithm co-design solution incorporating the device constraints (low TMR, low dynamic range) and circuit constraints (non-linearity, process variations, area limitations based on the extensive monte-carlo simulations and parameters utilizing GF22nm FD-SOI technology node) into our algorithmic framework, resulting in an average of 45.3% reduction in backend-processor energy consumption on the NMNIST, CIFAR10-DVS, IBM DVS128-Gesture dataset, albeit with a 0.28%, 1.55%, and 1.23% in test accuracy drop from the baseline, respectively.

To the best of our knowledge, this is the first work to present a pathway towards attaining *high-accuracy and programmability on complex neuromorphic datasets* using in-pixel processing for NVS cameras leveraging a CMOS+X solution.

2 DEVICE PRELIMINARIES AND MODELING

Figure 1(a) depicts the structure of a spin-orbit-torque (SOT)-based MDWMTJ. It consists of a thin insulating oxide layer between two ferromagnetic (FM) layers: a pinned layer (PL) with fixed magnetization and a free layer (FL) with parallel (P) and antiparallel (AP) magnetic domains. The AP domain has higher resistance due to the tunnel magnetoresistance (TMR) effect. A heavy metal (HM) layer below the FL influences the domain wall (DW) position through current-induced SOT and Dzyaloshinskii-Moriya interaction (DMI) [25, 34, 35]. Unlike conventional MTJs, the MDW-MTJ exhibits continuous resistance states based on the DW position (q), allowing for multiple resistance levels, as experimentally demonstrated [20].

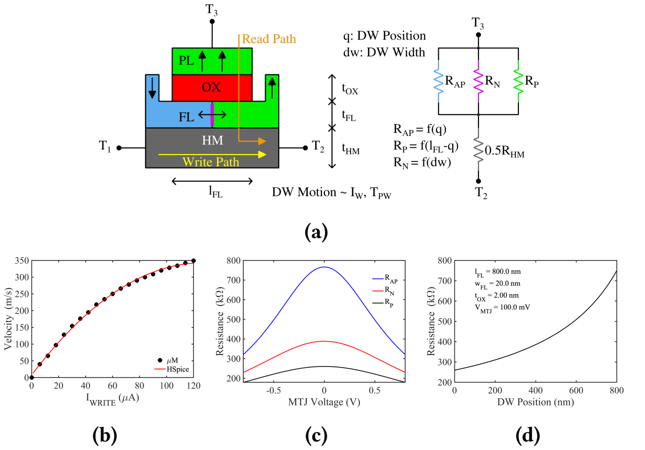


Figure 1: (a) Device structures for SOT-based MDWMTJ. l_{FL} t_{FL} **denote the length and thickness of the FL,** t_{OX} **and** t_{HM} **represent the thickness of the oxide and HM layer, respectively.** R_P, R_{AP}, **and** R_N **denote the MTJ resistance of the parallel, and-parallel and perpendicular state. (b) DW velocity as a function of the write current. (c) MTJ resistance as a function of applied voltage. (d) Effective MDWMTJ resistance as a function of the DW position.**

In the SOT-based MDWMTJ, the read and write paths are decoupled, allowing independent optimization. The DW position is programmed by applying a current pulse through the HM between the terminal T1 and T2 (Figure 1(a)), where the DW motion is proportional to pulse amplitude and duration. Our approach utilizes the T1 to T2 directed (positive direction) current flow for weighted accumulation, a configuration experimentally demonstrated in [23]. A long-duration current from T2 to T1 (negative direction) is applied to reset the DW position (q), which shifts the DW to the leftmost position (q = 0). The MDWMTJ resistance (between T3 and T2) is non-linearly dependent on DW position and applied voltage. At q = 0 and q = l_{FL}, MDWMTJ exhibits the lowest (R_P) and highest (R_{AP}) resistance, respectively. For other DW positions (q), the resistance non-linearly varies between R_P and R_{AP}.

We developed a compact Verilog-A model of the MDWMTJ for HSpice simulations, benchmarking it with MuMax3 μM results utilizing experimental device parameters from [34]. The Verilog-A model captures DW velocity as a function of the write current utilizing a 2nd-order polynomial fitting function (Figure 1(b)), demonstrating 1.06% of RMSE error. Using a resistance model that has been benchmarked to a non-equilibrium Green's function (NEGF)-based framework from [11], we simulated MTJ resistance across different applied voltages (Figure 1(c)). To determine the resistance of an MTJ in an FM with a domain wall between two oppositely polarized domains, the NEGF-based simulator is adapted to account for a parallel connection of three separate MTJs (parallel, anti-parallel, and perpendicular) [31]. The parallel and anti-parallel resistance varies as a function of the DW position, while the perpendicular resistance depends on the DW width. The effective MDWMTJ resistance, including the HM resistance as a function of

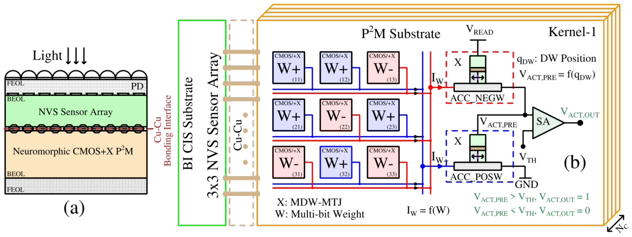


Figure 2: (a) The representative 3D heterogeneously integrated CMOS+X P^{2}M architecture utilizing Cu-Cu hybrid bonding, where the top die is backside illuminated CIS substrate, and the bottom die consists of P^{2}M compute elements. (b) A computing architecture of the MAC and thresholding operation considering a kernel size of 3×3.

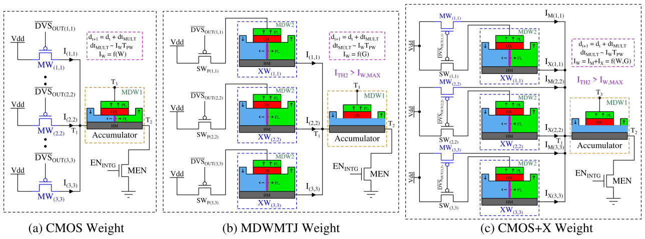
the DW position, is shown in Figure 1(d). The compact 3-terminal MDWMTJ Verilog-A model capturing the DW dynamics and non-linear voltage-dependent resistance is utilized in our analysis.

3 PROPOSED NEUROMORPHIC CMOS+X P^{2}M HARDWARE ARCHITECTURE

This section introduces the hardware innovations and implementation of our proposed CMOS+X spiking CNN architecture. The in-situ P^{2}M hardware can be heterogeneously 3D integrated [16, 26] where the top die accommodates the NVS array, and the bottom die contains vertically aligned MAC units per the spiking CNN filter inputs (Figure 2). The sensor array generates an ON (OFF) event spike when the contrast detected by the neuromorphic pixel increases (decreases) by a certain threshold. The neuromorphic CMOS+X P^{2}M core on the bottom die receives event spikes as inputs from the top die through hybrid Cu-Cu bonding. The P^{2}M core consists of analog weights and accumulators to perform the MAC computations. The current represents the weight values encoded as device parameters, such as the width of CMOS transistors, or the resistance of the MDWMTJ, or both. The input-current-modulated DW motion of the MDWMTJ has been utilized as the accumulator.

Spiking CNN requires asynchronous MAC computation and synchronous thresholding after a fixed integration time. Input event spikes trigger asynchronous write currents through the accumulator MDWMTJ, dependent on weight values. For large (small) weight values, large (small) write current (I_W) flows through the HM, resulting in large (small) DW motion from its previous state (position). As the input spikes are binary, DW motion represents the accumulated MAC output, with separate MDWMTJs for positive (ACC_POSW) and negative (ACC_NEGW) weights. After the integration time, the pre-activation voltage ($V_{ACT,PRE}$), non-linearly proportional to DW position differences of the positive and negative accumulators, is generated. Finally, a thresholding circuit compares the pre-activation voltage with the threshold, producing the output activation spike for the next layer if it exceeds the threshold. A tunable threshold per neuron is achieved using a series divider of two programmable MDWMTJs, which will be utilized in the algorithmic optimization, along with the programmable weights for catering to various applications.

The spiking CNN requires multiple channels in the first layer for improved accuracy. Each channel operates independently with its weights and accumulators, performing asynchronous MAC and synchronous thresholding in parallel. After the thresholding, one channel is activated at a time, and the output activations of the different channels are read sequentially utilizing the asynchronous Address-Event Representation (AER) read scheme [4, 22] similar to the neuromorphic vision sensors. The output activation map is determined by kernel size and stride; hence, it is smaller than the raw sensor array. In addition, it eliminates the need for an extra bit to indicate event polarity (ON or OFF); hence, it reduces the number of required communicated off-chip address bits. Upon spike generation, a reset phase is executed for both positive and negative accumulators in the channel, involving a reset current flow in the negative direction to move the DW to the leftmost position.



(a) CMOS Weight (b) MDWMTJ Weight (c) CMOS+X Weight

Figure 3: Embedded in-situ multi-bit (a) CMOS-based, (b) MDWMTJ-based, and (c) CMOS+X-based weight implementation. Transistors $MW_{x,y}$, MDWMTJs $XW_{x,y}$ and transistors $MW_{x,y}$ + MDWMTJs $XW_{x,y}$ represent the weights in (a), (b), and (c), respectively, where, (x,y) = (1,1), (1,2), ... (3,3), considering a kernel size of 3×3.

3.1 In-situ Multi-bit Weights

Figure 3 presents three configurations: CMOS, MDWMTJ, and Hybrid CMOS+X weights to perform the MAC operation. In the CMOS-based implementation, weights (e.g., $MW_{(1,1)}$, $MW_{(2,2)}$, $MW_{(3,3)}$) are represented by the transistor’s width. The MDWMTJ-based configuration employs MDWMTJ as the weight (e.g., $XW_{(1,1)}$, $XW_{(2,2)}$, $XW_{(3,3)}$), and the resistance state (DW position) of the MDWMTJ dictates the weight value. The threshold current of the weight MDWMTJs needs to be higher ($I_{TH2} > I_{W,MAX}$) to prevent the weight current from triggering the writing process into the weight MDWMTJ during MAC operations. The thickness of the FL can modulate the threshold current of the MDWMTJ [38]. Though MDWMTJ-based configuration offers programmability, it exhibits a smaller dynamic range (due to low TMR [1], TMR = 200% used in this work) than CMOS-based weights, which are fixed during the fabrication steps. To overcome these limitations and achieve versatility, we propose a hybrid CMOS+X configuration, combining transistor width and MDWMTJ resistance for adjustable write currents. This hybrid approach provides a broad dynamic range akin to CMOS-based weights and modest tunability (30-40% for most weight values) like the MDW-based weights, requiring device-algorithm co-design optimization for effective implementation across various applications. Positive (negative) weights are connected to the positive (negative) accumulator MDWMTJs. A larger weight value corresponds to a wider transistor or/and a lower MDWMTJ resistance, resulting in a higher current flowing through the HM of the accumulator MDWMTJs during an event spike. The DW motion, triggered by the input spike, moves from left to right and accumulates the MAC

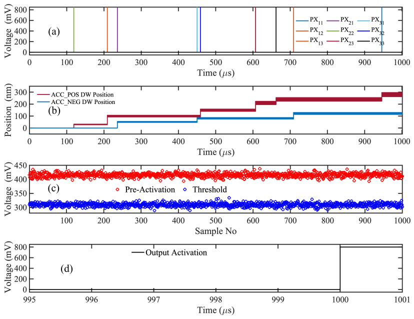


Figure 4: 1000 monte-carlo simulations of a random convolution operation with output activation spike.

results throughout the integration time. Due to non-volatility, the accumulator MDWMTJs retain the previous membrane potential (DW position) for subsequent integration times, a behavior important for achieving a high classification accuracy.

3.2 Analog Convolution Operation

Figure 4 presents an HSpice simulation of our hybrid CMOS+X approach based on asynchronous MAC computations and synchronous output activation using GF22nm FD-SOI technology. The simulation considers a 3×3 kernel size, random positive and negative weights, and random event timings and includes transistor 3-sigma variation, 10% Gaussian jitter in the write pulse, and 20% resistance variations in MDWMTJs. Subplot (b) illustrates the rightward movement of positive and negative accumulators during event spikes, as seen in subplot (a). Despite using a 1 ms integration time for efficient 1000 monte-carlo simulations, the MDWMTJ's non-volatility allows longer state retention, with cumulative CMOS leakage staying below the accumulator MDWMTJ's threshold current, preventing any DW movement. After 1 ms of integration time, the reasonable sense margin between the final pre-activation and threshold voltage for 1000 samples (subplot (c)) leads to an output activation spike for the next layer (subplot (d)).

3.3 Reprogrammability of Weight and Neuron's Threshold

In Figure 5, we showcase the reprogrammability of weights and threshold voltage in our hybrid CMOS+X configuration, where we have utilized both CMOS and MDWMTJ as weights. The blue and red shaded regions represent MAC and thresholding operations for initial and updated setups, while the green region indicates the programmability of weights and threshold voltage, along with accumulator reset. One positive and one negative weight were randomly selected for this test. The transistor width is fixed in the CMOS+X hybrid configuration, and MDWMTJ resistance can be tuned for weight modulation. Initially (blue-shaded region), the MDWMTJ of the positive (negative) weight is set to a higher (lower) resistance state by programming the DW position. Depending on input event spikes, the DW of the positive (ACC_POS DW in subplot (a)) and negative (ACC_NEG DW in subplot (a)) accumulators moves rightward from their reset state. From subplots (d) and (e), it can be observed that the pre-activation voltage is smaller than the threshold voltage for the initial weight setup, resulting in an output activation of 0. The reset and programmable features are then demonstrated. The accumulators are reset, and the DW positions of the positive and negative MDWMTJ weights are programmed to different states within the tunable range (subplot (c)). Additionally, the DW position of the reference voltage generation is programmed to a lower resistance state to achieve a smaller neuron threshold. Due to the updated weights and threshold voltage, the MAC and threshold results differ from the previous application. The difference between the positive and negative accumulator has increased, leading to a higher pre-activation voltage, surpassing the lowered threshold voltage, and resulting in an output activation spike.

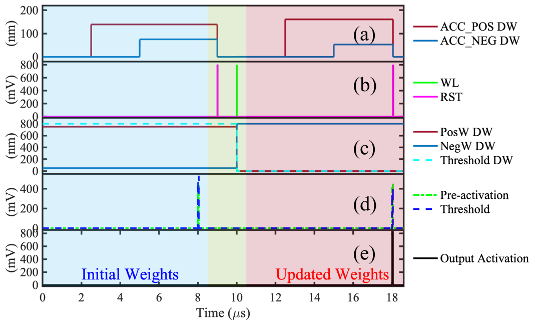
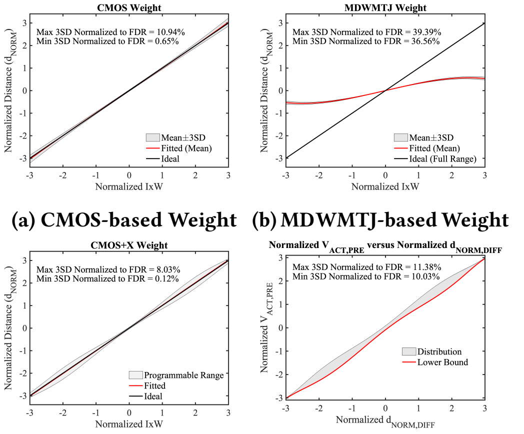


Figure 5: Weight and threshold voltage reprogrammability simulation for two different applications.

4 DEVICE-CIRCUIT-ALGORITHM CO-DESIGN

In this section, we detail the implementation of our algorithmic framework on the proposed neuromorphic CMOS+X P^2M architecture, considering both device and circuit constraints. We address the non-ideal, non-linear attributes of MDWMTJ and transistors, incorporating process variations to account for potential resistance and current deviations in our CMOS+X system. Through extensive HSpice simulations on the GF22nm FD-SOI node, we map and integrate our circuit characteristics into our spiking CNN framework, utilizing custom functions that adapt to circuit and device non-linearity, non-ideality, and variations.

MAC output Modeling: We encode spiking CNN filter weights as the current through the HM of accumulator MDWMTJs, adjusting the weight transistor's width or/and weight MDWMTJ's resistance (DW position). Domain wall motion depends on the write current pulse, exhibiting a non-linear relationship with the current (Figure 1(b)). Moreover, the non-linear and voltage-dependent write current by the transistor or/and MDWMTJ is also affected by process variations, fabrication uncertainty, and noise. Simulating for various input spikes and weight combinations, accounting for 3-sigma variation for GF22nm FD-SOI devices, 20% MDWMTJ resistance variation, and 10% clock jitter in the write pulse, our HSpice results for CMOS-based, MDWMTJ-based, and CMOS+X-based configurations (Figures 6(a), (b), and (c), respectively) are modeled using a behavioral curve-fitting function (f_1). Due to the low dynamic range (DR), the MDWMTJ-based weight exhibits higher worst-case variation (39.39%) compared to the CMOS (10.94%) and CMOS+X (8.03%) weights normalized to full DR. The CMOS+X configuration reports a 30-40% normalized programmable range for most weights.



(a) CMOS-based Weight **(b) MDWMTJ-based Weight**

(c) CMOS+X Weight **(d) Read Voltage**

Figure 6: Normalized DW position (d_{NORM}) versus the normalized input activation×weight characteristics for (a) CMOS, (b) MDWMTJ, and (c) CMOS+X weight configurations considering non-linearity and process variations. (d) Normalized pre-activation voltage ($V_{ACT,PRE}$) versus the difference in MAC results ($d_{NORM,DIFF}$) of the accumulators.

Pre-activation result modeling: After accumulating the DW motion for a fixed integration time, the pre-activation voltage is generated from the series division of the negative and positive accumulator MDWMTJs, where the subtracted outcome (positive MAC output - negative MAC output) non-linearly influences the output voltage of the series divider. Due to the non-linearity and voltage-dependent resistance of the MDWMTJ, identical differences in MAC results may generate different pre-activation voltages. Addressing this non-linearity through extensive HSpice simulations for various combinations (depicted in Figure 6(d)), we model the worst-case scenario (lower bound) that exhibits a maximum of 11.38% variations, utilizing a behavioral curve-fitting function (f_2) for the normalized pre-activation voltage versus the normalized distance difference of the accumulator MDWMTJs.

Circuit-Algorithm co-design optimization: Our algorithmic framework generates random Gaussian sample values based on mean and standard deviation results from HSpice simulations to address process variation effects. The accumulated DW position for each pixel's event spike is calculated using the f_1 function incorporating hardware's non-ideality, aggregated throughout the integration time, to determine MAC results for positive and negative weights. The normalized pre-activation voltage is computed through the f_2 function to incorporate the non-linearity and variations of the read circuits. The worst-case scenario (lower bound) is used for pre-activation voltage calculation to ensure accurate output spike generation. This algorithmic framework, employing custom functions f_1 and f_2, optimizes spiking CNN training for event-driven neuromorphic datasets. In addition, our algorithmic framework is also optimized utilizing 32 channels with stride 2 for the first layer, ensuring no area overhead for our P^2M core.

Device-Algorithm co-design optimization: Due to the limited TMR, our hybrid CMOS+X P^2M configuration offers a programmability range 30-40% from its median value. We employ weight tunability restrictions in our algorithmic framework to accommodate diverse applications, enabling our hardware to cater to various applications. Our approach involves initially training weights for one application and subsequently retraining them, considering the normalized weight tunability range derived from HSpice simulation results for another application. Additionally, we leverage kernel-dependent adjustments to neuron thresholds to optimize the algorithmic accuracy.

Table 1: Evaluation of our P^2M approach modeled using a custom first layer for NVS datasets, where the noise denotes the variation in the custom functions f_1 and f_2.

Dataset	Custom Func.	Noise (%)	Accuracy (%)
NMNIST	No (Baseline)	0	98.10
	Yes	0	98.04
	Yes	10	97.82
	Yes	20	97.71
	Yes	40	95.42
CIFAR10-DVS	No (Baseline)	0	80.72
	Yes	0	80.25
	Yes	10	79.17
	Yes	20	77.80
	Yes	40	65.58
DVS128-Gesture	No (Baseline)	0	97.22
	Yes	0	96.53
	Yes	10	95.99
	Yes	20	95.15
	Yes	40	87.25

5 EXPERIMENTAL RESULTS

In this work, we explore and benchmark our proposed neuromorphic CMOS+X P^2M solution utilizing the event-driven neuromorphic tasks, specifically classifying video samples captured by NVS cameras. The evaluation is conducted on three widely used neuromorphic benchmark datasets: NMNIST [28], CIFAR10-DVS [21], and IBM DVS128-Gesture [2]. The Spikingjelly package [10] is employed to process and integrate data into fixed time intervals, and a 9:1 train-validation split is applied for these datasets. The spiking CNN architecture consists of four convolutional layers, succeeded by two linear layers featuring 512 and 10 neurons, respectively. Following each convolutional layer, there is a sequence of a batch normalization layer, a spiking LIF layer, and a max pooling layer.

5.1 Classification Accuracy & Programmablity

Our spiking CNNs with in-pixel processing achieve test accuracy comparable to SNNs processed with traditional digital hardware (baseline), as demonstrated in Table 1. This success is attributed to the non-volatility of MDWMTJ and our algorithmic optimization considering variations. Additionally, our CMOS+X approach allows re-programmability, enabling post-fabrication tuning of the first layer weights in our P^2M hardware for diverse neuromorphic applications. As indicated in Table 2, fine-tuning, especially of the first layer, is essential for maintaining accuracy when transitioning from training on the DVS-CIFAR10 or IBM DVS128-Gesture datasets to other applications. The absence of this fine-tuning results in a significant accuracy drop.

Table 2: Efficacy of the reprogrammability of our proposed CMOS+X approach.

Train	Evaluate	Accuracy (%)		
		No Fine-tune	Except 1^{st} layer	Fine-tune
CIFAR10-DVS	DVS128-Gesture	9.12	94.96	96.51
DVS128-Gesture	CIFAR10-DVS	10.10	76.73	80.14

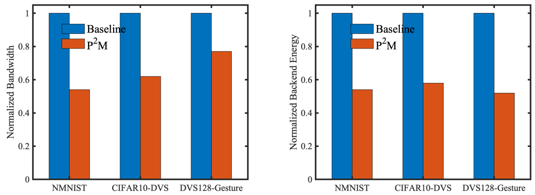


Figure 7: Normalized bandwidth (left-subplot) and backend energy (right-subplot) for the baseline and P^2M-enabled spiking CNN for NVS datasets.

5.2 Details of Bandwidth & Energy Savings

Fig. 7 demonstrates the bandwidth and energy savings with our in-pixel processing approach for the three neuromorphic datasets. We calculate bandwidth as the ratio of average output activation spikes to input event spikes per input sample. To reduce the bandwidth, we employ the bit-level ℓ_1 regularizer in the first layer, similar to [30]. The normalized bandwidth reductions of our approach compared to the baseline, where the input spikes are directly sent to the backend processor, are 0.54, 0.62, and 0.77 for NMNIST, DVS-CIFAR10, and IBM DVS128-Gesture, respectively. The backend compute energy consumption is normalized with respect to the conventional backend processing (i.e., digital implementation). Due to the in-pixel processing of the first spiking CNN layer in the analog domain, our approach yields an average (across 3 NVS datasets) of 45.3% lower backend energy. Note that the backend energy is determined by the number of spikes generated by each layer and the memory access of the weights and membrane potential, similar to [36].

6 CONCLUSION

We have proposed and implemented a novel CMOS+X processing-in-pixel-in-memory paradigm for neuromorphic event-based sensors. To our knowledge, this is the first CMOS+X implementation proposal for in-pixel processing focusing on neuromorphic vision sensor applications. Leveraging the non-volatility and programmable features of the MDWMTJ and the high dynamic range achieved by the CMOS transistors, our developed spiking CNN hardware can cater to various applications, yielding SOTA accuracy. In addition, we optimize and evaluate our system, incorporating device and circuit constraints into our algorithmic framework based on extensive HSpice simulations. Our neuromorphic CMOS+X P^2M-enabled spiking CNN model yields an accuracy of 97.82%, 79.17%, and 95.99% on the NMNIST, CIFAR10-DVS, IBM DVS128-Gesture datasets, respectively and achieved an average of 45.3% backend energy reduction compared to the conventional system.

ACKNOWLEDGMENTS

This work was funded in part by National Science Foundation (NSF).

REFERENCES

[1] Mahshid Alamdar et al. 2021. Domain wall-magnetic tunnel junction spin-orbit torque devices and circuits for in-memory computing. *APL* 118, 11 (2021).

[2] Arnon Amir et al. 2017. A Low Power, Fully Event-Based Gesture Recognition System. In *CVPR 2017*, Vol. 1. 7388–7397.

[3] Aayush Ankit et al. 2017. Resparc: A reconfigurable and energy-efficient architecture with memristive crossbars for deep spiking NN. In *DAC 2017*.

[4] Kwabena A Boahen. 2004. A burst-mode word-serial address-event link-I: Transmitter design. *IEEE TCAS-I* 51, 7 (2004), 1269–1280.

[5] Yang Chai. 2020. In-sensor computing for machine vision.

[6] Guang Chen et al. 2020. Event-based neuromorphic vision for autonomous driving: A paradigm shift for bio-inspired visual sensing and perception. *IEEE Signal Processing Magazine* 37, 4 (2020), 34–49.

[7] Gourav Datta et al. 2022. Can Deep Neural Networks be Converted to Ultra Low-Latency Spiking Neural Networks?. In *DATE 2022*, Vol. 1. 718–723.

[8] Gourav Datta et al. 2022. A processing-in-pixel-in-memory paradigm for resource-constrained TinyML applications. *Scientific Reports* 12 (2022).

[9] Ryoji Eki et al. 2021. A 1/2.3 inch 12.3 Mpixel with on-chip 4.97 TOPS/W CNN processor back-illuminated stacked CIS. In *ISSCC 2021*, Vol. 64. IEEE, 154–156.

[10] Wei Fang et al. 2020. SpikingJelly.

[11] Xuanyao Fong et al. 2011. KNACK: A hybrid spin-charge mixed-mode simulator for evaluating different genres of SOT MRAM bit-cells. In *2011 International Conference on Simulation of Semiconductor Processes and Devices*. IEEE, 51–54.

[12] Tzu-Hsiang Hsu et al. 2020. A 0.5-V real-time computational CIS with programmable kernel for feature extraction. *IEEE JSSC* 56, 5 (2020), 1588–1596.

[13] Tzu-Hsiang Hsu et al. 2022. A 0.8 V Intelligent Vision Sensor with Tiny CNN and Programmable Weights Using Mixed-Mode Processing-in-Sensor Technique for Image Classification. In *ISSCC 2022*, Vol. 65. IEEE, 1–3.

[14] Xuan Hu et al. 2019. SPICE-only model for spin-transfer torque domain wall MTJ logic. *IEEE TED* 66, 6 (2019), 2817–2821.

[15] S Ikeda et al. 2008. Tunnel magnetoresistance of 604% at 300K by suppression of Ta diffusion in CoFeB/ MgO/ CoFeB pseudo-spin-valves annealed at high temperature. *APL* 93, 8 (2008).

[16] Y Kagawa et al. 2020. Impacts of Misalignment on 1μm Pitch Cu-Cu Hybrid Bonding. In *IITC 2020*. IEEE, 148–150.

[17] Md Abdullah-Al Kaiser et al. 2023. Neuromorphic-P2M: processing-in-pixel-in-memory paradigm for neuromorphic image sensors. *Frontiers in Neuroinformatics* 17 (2023), 1144301.

[18] Martin Lefebvre et al. 2021. A 0.2-to-3.6 TOPS/W programmable convolutional imager soc with in-sensor current-domain ternary-weighted MAC operations for feature extraction and region-of-interest detection. In *ISSCC 2021*, Vol. 64. IEEE.

[19] Juan Antonio Leñero-Bardallo et al. 2011. A 3.6 μs Latency Asynchronous Frame-Free Event-Driven Dynamic-Vision-Sensor. *IEEE JSSC* 46, 6 (2011), 1443–1455.

[20] Thomas Leonard et al. 2022. Shape-Dependent Multi-Weight Magnetic Artificial Synapses for Neuromorphic Computing. *Advanced Electronic Materials* 8, 12 (2022), 2200563.

[21] Hongmin Li et al. 2017. CIFAR10-DVS: An Event-Stream Dataset for Object Classification. *Frontiers in Neuroscience* 11 (2017).

[22] Patrick Lichtsteiner et al. 2008. A 128x128 120 dB 15 μs latency asynchronous temporal contrast vision sensor. *IEEE JSSC* 43, 2 (2008), 566–576.

[23] Shijiang Luo et al. 2021. Integrator based on current-controlled magnetic domain wall. *APL* 118, 5 (2021).

[24] Ana I Maqueda et al. 2018. Event-based vision meets deep learning on steering prediction for self-driving cars. In *IEEE CVPR*. 5419–5427.

[25] Eduardo Martinez et al. 2014. Current-driven dynamics of Dzyaloshinskii domain walls in the presence of in-plane fields: Full micromagnetic and one-dimensional analysis. *Journal of Applied Physics* 115, 21 (2014).

[26] Tsukasa Miura et al. 2019. A 6.9 μm pixel-pitch 3D stacked global shutter CIS with 3M Cu-Cu connections. In *3DIC 2019*. IEEE, 1–2.

[27] Anh Nguyen et al. 2019. Real-time 6DOF pose relocalization for event cameras with stacked spatial LSTM networks. In *IEEE CVPR*. 0–0.

[28] Garrick Orchard et al. 2015. Converting Static Image Datasets to Spiking Neuromorphic Datasets Using Saccades. *Frontiers in Neuroscience* 9 (2015).

[29] Kaushik Roy et al. 2020. In-memory computing in emerging memory technologies for machine learning: An overview. In *DAC 2020*. IEEE, 1–6.

[30] Yusuke Sekikawa et al. 2023. Bit-Pruning: A Sparse Multiplication-Less Dot-Product. In *The Eleventh International Conference on Learning Representations*.

[31] Abhronil Sengupta et al. 2016. Proposal for an all-spin artificial neural network: Emulating neural and synaptic functionalities through domain wall motion in ferromagnets. *TBioCAS* 10, 6 (2016), 1152–1160.

[32] Ruibing Song et al. 2022. A reconfigurable convolution-in-pixel CIS architecture. *TCSVT* 32, 10 (2022), 7212–7225.

[33] Sepehr Tabrizchi et al. 2023. AppCiP: Energy-Efficient Approximate Convolution-in-Pixel Scheme for Neural Network Acceleration. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* 13, 1 (2023), 225–236.

[34] Chao Wang et al. 2020. Compact model of Dzyaloshinskii domain wall motion-based MTJ for spin neural networks. *IEEE TED* 67, 6 (2020), 2621–2626.

[35] Manman Wang et al. 2021. Compact model of domain wall MTJ driven by spin-orbit torque and Dzyaloshinskii–Moriya interaction. *IEEE Transactions on Magnetics* 58, 8 (2021), 1–5.

[36] R Yin et al. 2023. SATA: Sparsity-Aware Training Accelerator for Spiking Neural Networks. *TCAD* 42, 6 (2023), 1926–1938.

[37] Xueyong Zhang et al. 2022. A 915–1220 TOPS/W, 976–1301 GOPS Hybrid In-Memory Computing Based Always-On Image Processing for Neuromorphic Vision Sensors. *IEEE JSSC* 58, 3 (2022), 589–599.

[38] Daoqian Zhu et al. 2020. Threshold current density for perpendicular magnetization switching through SOT. *Physical Review Applied* 13, 4 (2020), 044078.