

Received 16 June 2026, accepted 25 June 2026, date of publication 3 July 2026, date of current version 3 August 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3709837

RESEARCH ARTICLE

Hybrid-SNN: CMOS 8T SRAM-Based Spiking Neural Network (SNN) Accelerator With Hybrid Leaky-and-Integrate Fire (LIF) Neuron for Optimized Energy Efficiency and Error Tolerance

HONGGU KIM^{ID}, YERIM AN^{ID}, JAEYOUN KIM^{ID}, AND YONG SHIM^{ID}, (Member, IEEE)

Department of Intelligent Semiconductor Engineering, Chung-Ang University, Dongjak-gu, Seoul 06974, Republic of Korea

Corresponding author: Yong Shim (yongshim@cau.ac.kr)

This work was supported in part by the Chung-Ang University Graduate Research Scholarship, in 2024; in part by the National Research Council of Science and Technology (NST) through Korean Government [Ministry of Science and ICT (MSIT)] under Grant GTL24042-000; and in part by Korea Institute of Science and Technology (KIST) Institutional Program under Grant 26E0031.

ABSTRACT Spiking Neural Networks (SNNs) are emerging as a promising paradigm in artificial intelligence, effectively mimicking the dynamic behavior of biological neural systems. SNN models typically employ Deterministic Leaky Integrate-and-Fire (D-LIF) or Stochastic Leaky Integrate-and-Fire (S-LIF) neurons, each optimized for specific performance goals. D-LIF neurons maximize energy efficiency through spike sparsity, while S-LIF neurons improve error tolerance. However, existing hardware lacks the capability to support both neuron types simultaneously. In this paper, we demonstrated Hybrid-SNN: CMOS implemented SNN accelerator with Hybrid LIF (H-LIF) neuron which is capable of operating as either D-LIF neuron or S-LIF neuron in a single hardware platform, enabling flexible deployment of various Hybrid-SNN configurations. Software-level simulations show that strategically distributing D-LIF and S-LIF neurons across specific layers enables the Hybrid-SNN to leverage the strengths of both neuron types: D-LIF neurons improve energy efficiency, while S-LIF neurons enhance error tolerance. This Hybrid configuration enhances both energy efficiency and computational robustness, outperforming single-mode neuron configurations. Fabricated using CMOS 180nm process technology, the Hybrid-SNN achieves an energy consumption of 1.62 pJ/SOP, offering a $1.59\times$ improvement over its Stochastic counterpart, and exhibits a low accuracy drop of 7.67% under a 5% random weight deviation scenario, achieving a $1.64\times$ improvement over its Deterministic counterpart.

INDEX TERMS Spiking neural network, compute-in-memory (CIM), leaky integrate-and-fire neuron, stochastic neuron.

I. INTRODUCTION

Spiking Neural Networks (SNNs) are emerging third generation of Artificial Intelligence (AI) algorithms, which incorporates real-time binary events for its information processing [1]. This event-driven representation more closely mimics the computing principles of human brains, hence enabling low power consumption by sparse and efficient information transfer and computing [2], [3], [4].

The associate editor coordinating the review of this manuscript and approving it for publication was Ludovico Minati^{ID}.

In sync with the development of SNN algorithms, spike-driven neuromorphic architectures are being rapidly developed to enable energy-efficient SNN implementation on hardware [5], [6], [7], [8], [9].

Besides, the generally adopted activation neuron in the SNN model is Deterministic Leaky Integrate-and-Fire (D-LIF) neuron, which continuously integrates the spike input stimuli in a form of membrane potential and fires spike output when the membrane potential exceeds the specified threshold value [1], [10], [11]. This ‘deterministic’ neuronal behavior yet does not fully embrace the realistic human

brain computing feature which incorporates ‘stochastic’ neuronal behavior featured by noise [12], [13]. Recent studies on SNN algorithms have begun to adopt noise induced stochasticity into SNN model primitives such as activation neurons [14], [15] and weight synapses [16]. One of the key motivations for introducing stochasticity into SNN models is to improve robustness against random error sources. By incorporating probabilistic behaviors into activation neurons or synaptic weights, the impact of random perturbations can be statistically mitigated, thereby suppressing accuracy degradation [17]. For this reason, hardware realization of stochastic activation neurons has been investigated recently, to facilitate stochastic computations in the SNN algorithms on hardware platform [18], [19], [20], [21], [22], [23].

However, Stochastic neuron tends to generate a higher number of spikes compared to D-LIF neuron, which decreases spike sparsity and subsequently degrade the energy efficiency [17]. As an example from our work, the stochastic neuron (Stochastic-LIF) utilizes thermal noise and the metastability of two cross-coupled inverters as intrinsic randomness sources. It can generate stochastic output spikes even in cases where a D-LIF neuron would not fire. This increases the average spike count, which in turn increases switching activity and reduces energy efficiency compared to the D-LIF neuron. For this reason, in scenarios where the energy efficiency is critical, D-LIF neurons are generally the preferred choice. Nonetheless, stochastic neurons offer significant benefits, particularly in applications where error tolerance is important. In such cases, stochastic neurons may be more suitable than D-LIF neurons. Consequently, the selection of the output neuron type should be based on the specific performance goals: Energy efficiency or Error tolerance.

In this work, we present a Hybrid SNN accelerator that incorporates a Hybrid Leaky Integrate-and-Fire (H-LIF) neuron architecture, supporting operation in both Deterministic and Stochastic Leaky Integrate-and-Fire (D-LIF and S-LIF) neuron modes within a single hardware platform, enabling flexible implementation of various Hybrid-SNN configurations. Software-level behavioral simulations confirm that, by strategically distributing D-LIF and S-LIF neurons across specific network layers (e.g., layers 1, 2, 4, 5, 6, and 8 configured as D-LIF neurons, and layers 3, 7, and 9 as S-LIF neurons in Spiking VGG-9 model), the hybrid-mode SNN model leverages the complementary benefits of each neuron type: D-LIF neurons contribute to enhanced energy efficiency, while S-LIF neurons enhance error tolerance.

This layer-wise hybrid neuron allocation strategy enables simultaneous optimization of energy efficiency and computational robustness, achieving superior performance compared to single-mode neuron implementations. Fabricated using CMOS 180nm process technology, the Hybrid-SNN achieves an energy consumption of 1.62 pJ/SOP, representing a $1.59\times$ improvement in energy efficiency compared to its Stochastic counterpart (Stochastic-SNN), and sustains a minimal accuracy drop of 7.67% under a 5% random weight deviation

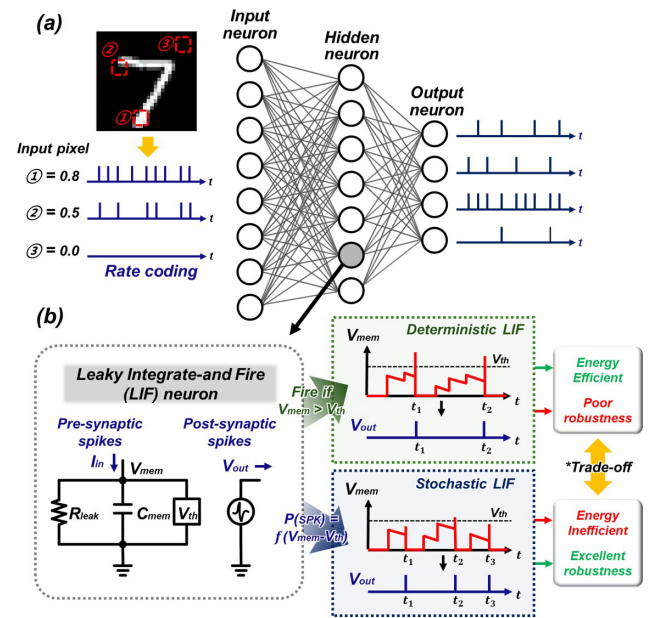


FIGURE 1. (a) Illustration of rate-coded Spiking Neural Network (SNN), (b) basic operational principles of Deterministic Leaky Integrate-and-Fire (D-LIF) and Stochastic LIF (S-LIF) neurons.

scenario, achieving a $1.64\times$ improvement in error tolerance over its Deterministic counterpart (Deterministic-SNN).

II. PRELIMINARIES

A. DATA ENCODING FOR SPIKING NEURAL NETWORK: RATE CODING

Rate coding is the most widely adopted data encoding scheme [5], [6]. As can be seen on Fig. 1(a), rate coding represents the data intensity by the number of spikes in the Poisson spike trains, where the input data intensity determines the spike-generation probability. Therefore, a larger data intensity statistically results in more frequent spike generation within a given spike interval, and the mean spike-generation rate converges to the original input data intensity over time. For instance, when an input = 0.5 (corresponding to a 50% spike-generation rate) is applied over 1,000 spike iterations, approximately 500 spike events are statistically generated on average.

Based on this principle, rate coding offers a straightforward and intuitive encoding scheme that is well-suited for hardware implementation. Its simplicity aligns closely with traditional digital signal processing methods, making it relatively easy to integrate into hardware systems. However, the accurate implementation of Poisson spike trains on silicon-based hardware presents significant challenges, which remains an active area of research in neuromorphic engineering [24].

Temporal coding schemes, such as Time-to-First-Spike (TTFS), are also widely considered efficient encoding methods in SNN models. However, temporal coding is highly sensitive to random perturbations [17]. Since information is encoded in the precise timing of spike events, small

timing deviations can directly degrade the inference accuracy. In contrast, rate coding averages spike activity over multiple time iterations, statistically suppressing the impact of the random errors. As a result, it is more robust to random perturbations. Considering the difference in noise sensitivity and the error tolerance, rate coding is adopted in this work.

B. DETERMINISTIC LIF (D-LIF) NEURON AND STOCHASTIC LIF (S-LIF) NEURON

Fig. 1(b) illustrates the fundamental operations of Deterministic-LIF (D-LIF) and Stochastic-LIF (S-LIF) neurons. Deterministic Leaky Integrate-and-Fire (D-LIF) neuron is the most widely utilized activation neuron in SNN models [1], [2], [3] and neuromorphic hardware [5], [6], which is a simplified mathematical model that captures some key aspects of neuronal behavior. D-LIF neuron integrates incoming presynaptic spikes over time, which is represented as an accumulation of membrane potential. When this membrane potential exceeds a predefined threshold, the neuron generates a postsynaptic spike and subsequently resets its membrane potential to the ground level.

Similar to the D-LIF neuron, the S-LIF neuron integrates presynaptic spikes over time in the form of membrane potential. However, S-LIF neuron is characterized by its probabilistic spiking mechanism. As the membrane potential approaches the threshold value, the ‘likelihood’ of generating a postsynaptic spike increases. Consequently, the S-LIF neuron can generate postsynaptic spikes even if the membrane potential has not reached the threshold, potentially resulting in a higher spike generation rate compared to the D-LIF neuron.

Consequently, the Deterministic Leaky Integrate-and-Fire (D-LIF) neuron exhibits superior energy efficiency compared to the Stochastic Leaky Integrate-and-Fire (S-LIF) neuron, primarily due to its reduced spike generation rate. Yet, the inherent uncertainty of the S-LIF neuron increases its resilience to random error sources, including signal noise and hardware non-idealities, yielding greater error tolerance than the D-LIF neuron [17]. In summary, as shown in the right side of 1(b), the D-LIF and S-LIF neurons present a ‘trade-off’ between ‘energy efficiency’ and ‘error tolerance’.

III. OVERALL ARCHITECTURE

Fig. 2 shows the overall architecture of proposed Hybrid-SNN accelerator. The Stochastic bit, utilized in the 64 input neurons, is a latch-based Poisson spike train generator with digital code inputs [24], driving the Source-Lines (SLs) of an 8T SRAM array and controlling two series-connected NMOS transistors within each bitcell. 8T SRAM array consists of 16 banks, where one consists of 64×4 8T SRAM based sub-block. In one row of one bank, four 8T SRAM cells constitute a 3-bit signed weight as was demonstrated in [25]. Specifically, the four 8T SRAM cells are divided into a positive-weight group and a negative-weight group. Each group consists of two 8T SRAM cells that represent the weight magnitude using $1\times$ and $2\times$ transistor width scaling,

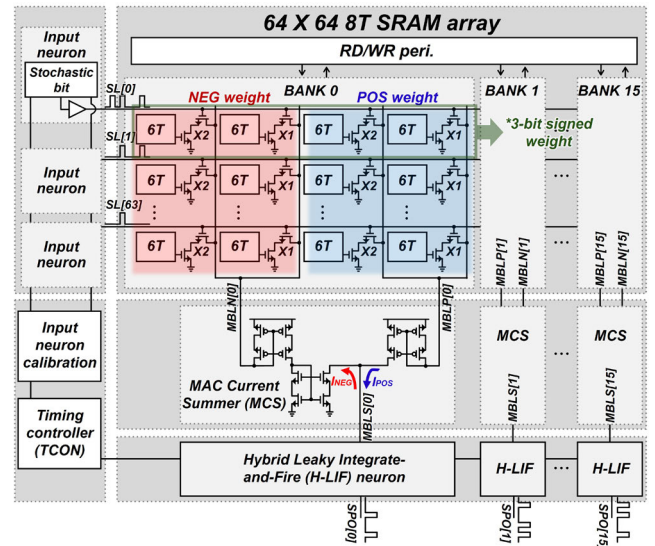


FIGURE 2. Overall architecture of the proposed Hybrid-SNN.

generating weighted current levels. The sign of the weight is determined by activating either the positive or negative group during the MAC operation. For example, to represent a weight of -2 , only the $2\times$ branch of the negative-weight group stores ‘1’, while all other cells store ‘0’. In contrast, to represent a weight of $+3$, both the $1\times$ and $2\times$ branches in the positive group store ‘1’, and the 8T SRAM cells in negative group store ‘0’.

The Poisson distributed voltage pulse trains that are generated at the input neurons are multiplied with the weight bits stored in 8T SRAM cells in the form of current flowing through the two series-connected NMOS transistors, and the resulting multiplied currents are accumulated on positive MAC bitline (MBLP) and negative MAC bitline (MBLN) separately, depending on whether the corresponding weight is positive or negative. The accumulated currents on MBLP and MBLN are then summed at MAC Current Summer (MCS) stage using cascode current mirrors. Currents on MBLP representing positive MAC values are converted to pull-up currents (I_{POS}) and currents on MBLN representing negative MAC values are converted to pull-down currents (I_{NEG}), and these two currents are merged at summed MAC bitline (MBLS) and are delivered to Hybrid LIF (H-LIF) neuron. Lastly, H-LIF neuron generates output spike train based on the current accumulations from MBLS lines.

A. RATE-CODED INPUT SPIKE GENERATOR: STOCHASTIC BIT

Fig. 3(a) shows the schematic of Stochastic bit developed from the previous work [24]. By leveraging thermal noise and the meta-stability of two cross-coupled inverters as sources of randomness, Stochastic bit generates Poisson-distributed spike trains by repeating the two operational steps: 1) Equalization (EQ) and 2) Evaluation (Eval), as shown in Fig. 3(b). In the EQ phase, nodes A and B of the

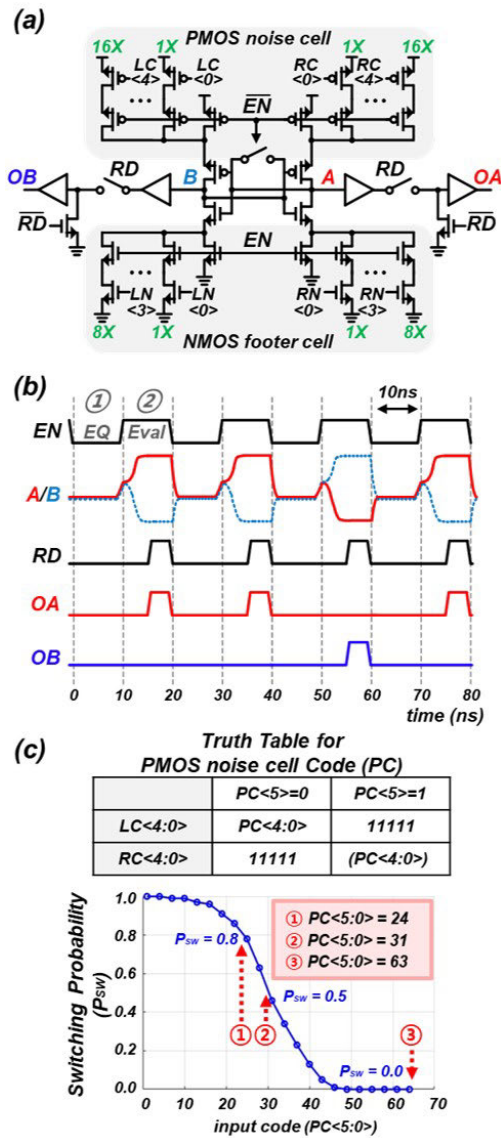


FIGURE 3. Stochastic bit [24] (a) Schematic, (b) operational timing diagram and (c) Switching probability (P_{sw}) and illustration of the Look-Up Table (LUT)-based P_{sw} mapping method.

cross-coupled inverters are equalized to the same voltage potential. In the subsequent Eval phase, nodes A and B are probabilistically driven to either 1/0 or 0/1, thereby generating the desired random spiking pattern.

The resulting A and B node voltages are subsequently buffered to output nodes OA and OB while the RD signal remains high. The OA output then controls the switching of the source line (SL) in the 8T SRAM array, as shown in Fig. 2, thereby enabling spiking current-based MAC operations. Although only OA is utilized for the subsequent MAC operation and OB remains unused, the complementary OB path is also implemented to preserve the structural symmetry between the left and right branches of the Stochastic bit, including the output buffer stages. This symmetric implementation helps minimize mismatch-induced offset

caused by parasitic RC imbalance, thereby maximizing the accuracy of stochastic spike generation.

In Fig. 3(a), the PMOS noise cell controls the asymmetric pull-up strength on the left and right sides of the cross-coupled inverter, thereby modulating the spike generation rate. Fig. 3(c) depicts the spike generation rate across the entire PMOS noise cell code (PC) range, exhibiting a sigmoidal curve based on the input code. The NMOS footer cell codes (LNs and RNs) are initially set to zero and are only activated as needed to calibrate current strength imbalances between the left and right current branches by introducing an additional asymmetric discharge current [26].

Additionally, the firing rate of the stochastic-bit spike train is determined based on a pre-characterized Look-Up Table (LUT) of the PMOS noise cell code (PC<5:0>). The LUT records the output spike probability corresponding to each PC code across the entire input range and is constructed from the estimated spike probability transfer curve of the stochastic bit. To generate the spike trains with a specific firing rate proportional to the input pixel intensity, the corresponding PC code in the LUT is selected and directly applied to the Stochastic bit. For instance, as shown in Fig. 3(c), PC codes of 24, 31, and 63 generate firing probabilities of approximately 0.8, 0.5, and 0.0, respectively.

B. HYBRID LEAKY-INTEGRATE AND FIRE (H-LIF) NEURON: DETERMINISTIC LIF (D-LIF) MODE AND STOCHASTIC LIF (S-LIF) MODE

Fig. 4(a) presents the schematic of the proposed Hybrid LIF (H-LIF) Neuron, whereas Fig. 4(b) illustrates the schematic of the Analog Activation Modules (AAM) employed within the H-LIF neuron. The passive circuit components, consisting of C_{mem} (4.5 pF) and R_{leak} (10.8 k Ω), continuously perform the Leaky-Integrate operation based on the pre-synaptic spikes delivered via the MAC current (I_{MAC}). Subsequently, the AAM evaluates whether to generate an output spike based on the accumulated membrane potential (V_{IN}). If an output spike is generated, then the D Flip-Flop latches the output spike and resets the membrane potential (V_{IN}) to zero by turning on the reset transistor (M_{RST}).

As shown in Fig. 4(b), The AAM comprises two cross-coupled inverters integrated with PMOS noise cell and Gain cell which regulate the pull-up strength of the respective current branches. The PMOS transistors in the PMOS noise cell, controlled by ENB<4:0> codes, generate an asymmetric ‘weak’ pull-up noise current to the two cross-coupled inverters through the top ports, enabling ‘stochastic’ output spike generation based on the difference between input voltage (V_{IN}) and threshold voltage (V_{th}). Conversely, the PMOS transistors in the Gain cell, controlled by ENB<5> code, deliver a ‘strong’ pull-up current to the two cross-coupled inverters via the middle port, ensuring ‘deterministic’ behavior as a comparator.

Note that the operation of the AAM is controlled by the SA_ENB signal, which determines the (1) reset and (2) enable phases of the module. When SA_ENB = 1

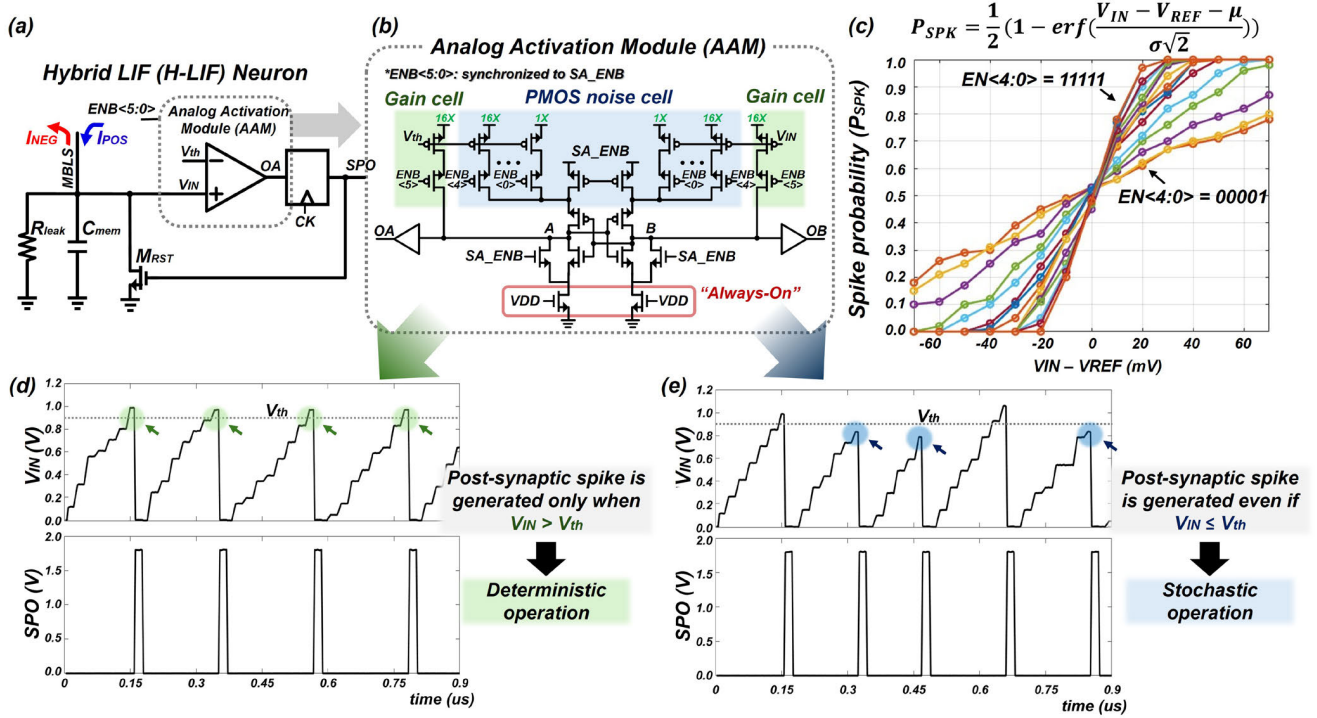


FIGURE 4. (a) The schematic of the proposed Hybrid LIF (H-LIF) neuron and (b) the schematic of Analog Activation Modules (AAM) utilized in H-LIF neuron, (c) simulated AAM spike probability versus $(V_{IN} - V_{th})$ with different PMOS noise cell codes ($EN<4:0>$), simulation result of H-LIF neuron operation with (d) D-LIF mode configuration and (e) S-LIF mode configuration.

(reset phase), all $ENB<5:0>$ codes are synchronously maintained at logic '1', thereby deactivating all PMOS transistors in the PMOS noise cell and Gain cell. At the same time, the NMOS reset transistors connected to nodes A and B are turned on, discharging both nodes to 0 V. Subsequently, when $SA_ENB = 0$ (enable phase), the reset NMOS transistors are turned off, and the $ENB<5:0>$ signals begin operating synchronously according to the programmed code values, where selected ENB signals are pulled low to activate the corresponding PMOS branches. Based on the $ENB<5:0>$ signals synchronized with the SA_ENB signal, the AAM operates either in D-LIF mode or S-LIF mode depending on the configured code values.

When D-LIF mode is configured on H-LIF neuron, the $ENB<5:0>$ codes are all set to low (turning on all PMOS transistors in PMOS noise cell and Gain cell) during the enable phase, ensuring the deterministic operation as a comparator. With this configuration on AAM, the output spike is generated only if the membrane potential (V_{IN}) exceeds the threshold voltage (V_{th}). On the contrary, when S-LIF mode is configured on H-LIF neuron, $ENB<5>$ is always set to high to disable the deterministic operation, and the impacts of the thermal noise and the meta-stability of two cross-coupled inverters become dominant. In this mode, only the weak pull-up strength of the PMOS noise cell is controlled via $ENB<4:0>$, realizing the stochastic spike generation based on the difference between input voltage (V_{IN}) and threshold voltage (V_{th}).

In addition, the "always-on" NMOS transistors shown in Fig. 4(b) are intentionally utilized to enhance the influence of thermal noise during S-LIF mode operation. The thermal noise characteristic of the cross-coupled inverter, expressed as $\bar{v}_n^2 = 4kT\gamma/g_m$, is inversely proportional to the transconductance (g_m) of the left and right current branches. Therefore, the introduction of the "always-on" transistors reduces the effective g_m , thereby increasing the impact of thermal noise and enhancing the stochastic switching behavior of the AAM. Although these transistors can introduce additional offset, it can be compensated using the calibration technique proposed in [26], which employs additional NMOS calibration transistors to compensate for the offset. On the other hand, during D-LIF mode operation, the Gain cell significantly increases the effective g_m , thereby suppressing the influence of thermal noise and enabling deterministic operation.

Fig. 4(c) shows the graph and the formula of the spike probability of AAM when it is configured to operate on S-LIF mode. The spike probability exhibits sigmoid curve based on the input differential voltage ($V_{IN} - V_{th}$), where the spike probability can be fitted to normal CDF formulation as was demonstrated in [27]. There are two more parameters that determine the shape of the sigmoid curve, the sigma (σ) that determines the slope of the sigmoid curve and the offset (μ) in two cross-coupled inverters. σ is controlled by PMOS noise cell codes as can be seen on Fig. 4(c). Lower $ENB<4:0>$ code values (activating more PMOS transistors) result

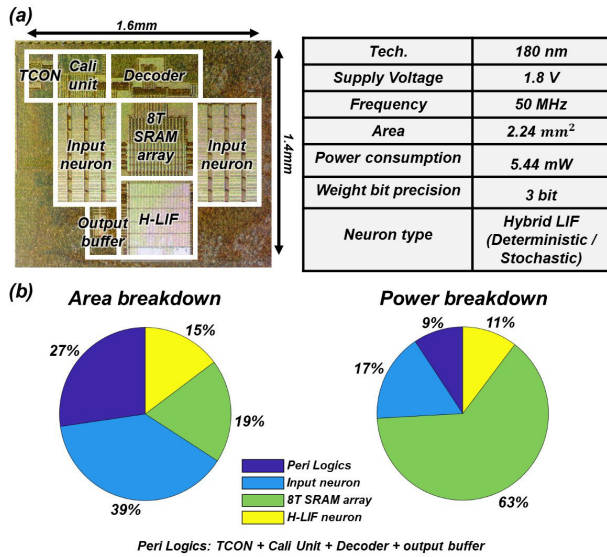


FIGURE 5. (a) Hybrid-SNN chip micrograph and the key specifications summary and (b) area and power breakdown.

in a smaller σ , leading to a steeper slope of the sigmoid curve, and vice versa. Moreover, the graph in Fig. 4(c) assumes that there is no offset in the two cross-coupled inverters ($\mu = 0$). But practically there exists a random offset in the two cross-coupled inverters due to device mismatch, and it might lead to horizontal offset in the sigmoid curve of spike probability.

Fig. 4(d) illustrates the simulated operation of the H-LIF neuron in D-LIF mode. The MAC current (I_{MAC}) is integrated in the form of the membrane potential (V_{IN}) over multiple time steps. Note that H-LIF neuron generates output spike (SPO) only if the membrane potential (V_{IN}) exceeds the threshold voltage (V_{th}), and V_{IN} is reset to zero. Meanwhile, Fig. 4(e) shows the simulated H-LIF neuron operation when S-LIF mode is configured. Like D-LIF mode operation, the MAC current (I_{MAC}) is continuously integrated in the form of membrane potential (V_{IN}). However, unlike the D-LIF mode, the S-LIF mode introduces a stochastic mechanism allowing the neuron to generate an output spike even if the membrane potential (V_{IN}) does not exceed the threshold voltage (V_{th}). Consequently, as shown in Fig. 4(e), the output spike (SPO) is generated at some point even if the membrane potential (V_{IN}) has not yet reached the threshold voltage (V_{th}). This stochastic operation results in a higher likelihood of spike generation in the S-LIF mode compared to the D-LIF mode.

IV. PERFORMANCE ANALYSIS

A. MEASUREMENT RESULTS

Fig. 5(a) shows the prototype chip that has been fabricated using the TSMC 180nm process. Additionally, the key performance metrics are detailed in the chip summary table depicted in Fig. 5(a). the total area of fabricated chip is 2.24 mm², and the entire system operates at a total power consumption of 5.44 mW, with a supply voltage

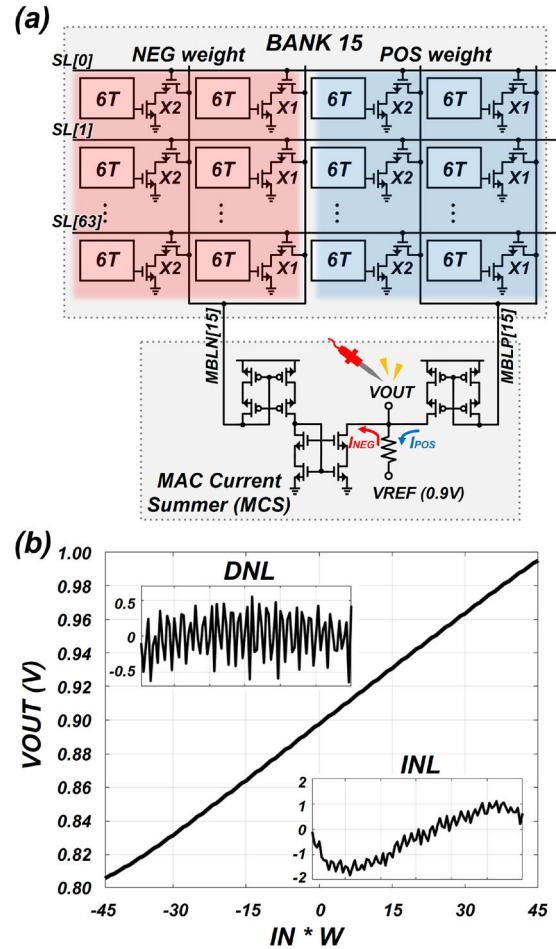


FIGURE 6. (a) Test setup for measuring the static MAC linearity in a single bank of the 8T SRAM array and (b) measured static MAC operation transfer curve along with the corresponding DNL and INL error plots.

of 1.8 V and a main clock frequency of 50 MHz. Fig. 5(b) presents the breakdown of both area and power consumption. In particular, the 64 × 64 8T SRAM array and the 64 input neurons occupy the largest portions of the chip area due to the expansion of each unit in the form of a uniform array. Furthermore, the 8T SRAM array accounts for the majority of power consumption due to the simultaneous discharge current across the entire array during MAC operation.

To evaluate the current-based MAC operation functionality of a single bank within the 8T SRAM array, we constructed the experimental setup as illustrated in Fig. 6(a). A 480-ohm polyresistor (R_{SS}) is directly connected to the VOUT node, where the total MAC currents are summed, to measure the MAC linearity in voltage form. Fig. 6(b) presents the transfer curve of the MAC operation output voltage (VOUT) and the corresponding calculated differential nonlinearity (DNL) and integral nonlinearity (INL) errors. The results are obtained by fixing all input activations to high (VDD) and sweeping weight codes from −45 to 45. The results demonstrate high linearity of the current-based MAC operation, with the

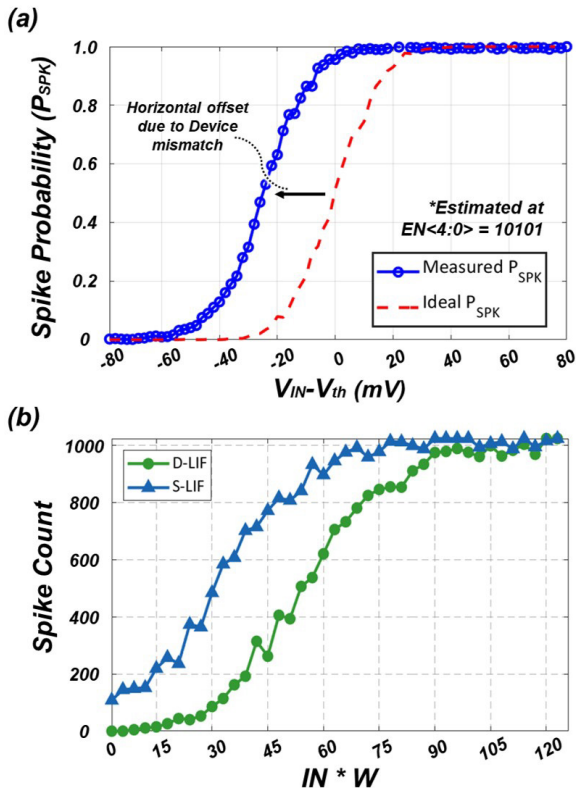


FIGURE 7. Measured spike probability (P_{SPK}) transfer curve of the Analog Activation Module (AAM) in the Hybrid LIF neuron with an EN code of 21 and (b) Measured spike count of the proposed Hybrid LIF neuron configured in D-LIF mode (green) and S-LIF mode (blue) over 1024 time steps.

worst-case DNL and INL errors of only -0.64 LSB and -2.08 LSB, respectively.

Fig. 7(a) shows the measurement result of the Analog Activation Module (AAM) in the Hybrid-LIF neuron, where it is configured to operate as the Stochastic LIF neuron mode with the PMOS noise cell code ($EN<4:0>$) of 21. The spike transfer curve exhibits a well-controlled sigmoid shape with respect to the differential input voltage ($V_{IN} - V_{th}$), though it shows a leftward horizontal offset. This horizontal offset is primarily due to device mismatches in the cross-coupled inverters within the AAM, which arise from random fluctuations in the CMOS fabrication process. To correct this offset, the calibration technique proposed in [26] can be applied. This method compensates for the current imbalance between the two branches of the cross-coupled inverters by introducing additional current through small calibration transistors. By cancelling out the current mismatch between the two branches, the transfer curve is aligned to the desired mid-crossing point, thereby mitigating the impact of fabrication-induced mismatches.

Fig. 7(b) illustrates the measurement results of the LIF operation for (1) the D-LIF neuron mode (green) and (2) the S-LIF neuron mode (blue), with an increasing MAC current pulse generated from the 8T SRAM array.

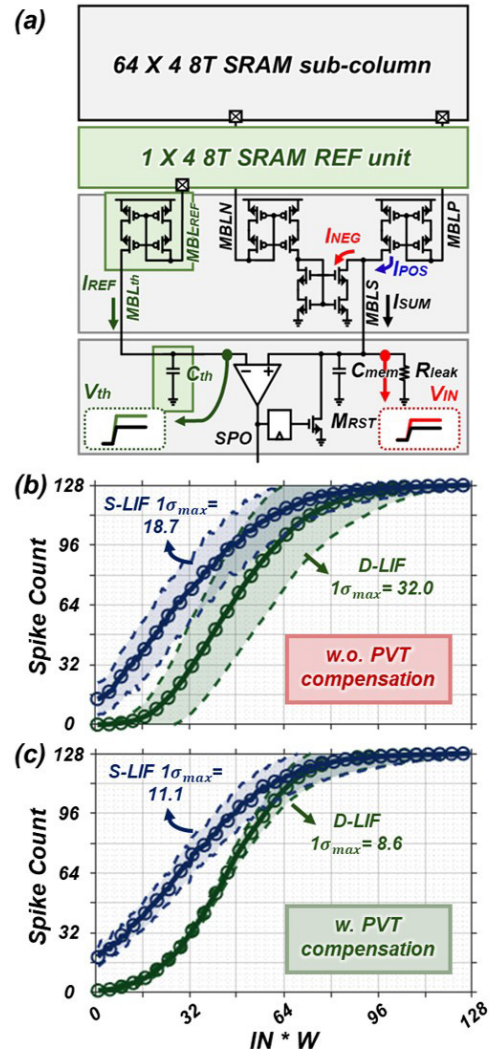


FIGURE 8. (a) Illustration of PVT compensation structure utilizing reference 8T SRAM unit for PVT tracking, 27 PVT corner simulated P_{SPK} transfer curves (b) without the PVT compensation structure and (c) with the PVT compensation structure.

The increment in MAC current is achieved by fixing the input neurons to produce spike pulses with a probability of 1, while monotonically increasing the 8T SRAM weight synapses in steps of 3. Since the S-LIF neuron mode can generate spikes even when the membrane potential does not reach the threshold voltage (V_{th}), it exhibits a higher likelihood of spike generation compared to the D-LIF neuron mode.

B. 27 PVT CORNER SIMULATION AND COMPENSATION TECHNIQUES

Since the proposed Hybrid-SNN architecture relies on analog current accumulation in the 8T SRAM array, it is sensitive to process, voltage, and temperature (PVT) variations, which can distort the subsequent spike probability (P_{SPK}) transfer curves. To investigate this effect, we performed simulations across 27 PVT corners and

developed an applicable compensation technique that effectively mitigates PVT-induced deviations and restores the original P_{SPK} transfer curves.

Fig. 8(a) illustrates the applicable PVT compensation technique, which introduces an additional 1×4 8T SRAM reference unit, an extra pair of PMOS current mirrors, and an additional capacitor (C_{th}) for V_{th} voltage sampling in each 64×4 8T SRAM sub-column. Using these components, the reference SRAM unit generates a reference current (I_{REF}) through MBL_{th} that reflects the current PVT condition at the beginning of each spiking inference event. This reference current is sampled onto C_{th} only once at the start of the inference process, generating a V_{th} voltage that tracks the corresponding PVT variation. As a result, the generated V_{th} can compensate for PVT-induced shifts in the P_{SPK} transfer characteristics.

Specifically, if the output summed current ($I_{SUM} = I_{POS} + I_{NEG}$) increases relative to the nominal operating condition due to a particular PVT corner (e.g., lower temperature, higher V_{DD} , or the FF process corner), the reference current (I_{REF}) also increases proportionally. This is because I_{REF} is generated from a reference 8T SRAM unit that has an identical structure to the computing 8T SRAM sub-column and is therefore subject to the same PVT variations. As a result, the final spike count can be effectively compensated across different PVT corners.

Figs. 8(b) and (c) present the simulation results across 27 PVT corners, where the process corners correspond to FF, TT, and SS, the supply voltage conditions are 2.0 V, 1.8 V, and 1.6 V, and the temperature conditions are set to -30°C , 30°C , and 90°C , respectively. Without PVT compensation, a fixed threshold voltage (V_{th}) is applied using an external voltage source, whereas with PVT compensation, the proposed scheme illustrated in Fig. 8(a) is adopted.

Without compensation (Fig. 8(b)), the 1σ variation of spike count is estimated as 18.7 and 32.0 for S-LIF and D-LIF modes, respectively. After applying the PVT compensation scheme, the corresponding 1σ values are reduced to 11.1 and 8.6 (Fig. 8(c)). It is important to note that the S-LIF operation inherently exhibits stochastic spiking behavior, which results in a relatively larger 1σ variation even under compensation. Interestingly, this inherent stochasticity contributes to improved robustness against PVT variations. In particular, without compensation, the 1σ variation increases only up to 18.7 in the S-LIF mode, whereas it increases significantly up to 32.0 in the D-LIF mode, indicating that the S-LIF dynamics are resilient against PVT-induced variations.

C. SOFTWARE-LEVEL ANALYSIS OF THREE DISTINCT SNN MODELS: DETERMINISTIC, STOCHASTIC AND THE PROPOSED HYBRID LIF NEURON IMPLEMENTATIONS

To evaluate the performance enhancement achieved through the layer-wise hybrid configuration of the proposed H-SNN structure, we developed and analyzed three distinct Spiking VGG-9 models for CIFAR-10 dataset classification.

These models share an identical structure (e.g., 64 time steps, 3-bit weight kernels) as shown in Fig. 9(a), except for the neuron types. Fig. 9(b) details the mapping of LIF neuron types across the three models.

The Deterministic-SNN (D-SNN) model employs D-LIF neurons across all LIF layers, with no S-LIF neurons. Conversely, the Stochastic-SNN (S-SNN) model incorporates only S-LIF neurons throughout all LIF layers. Finally, the Hybrid-SNN (H-SNN) model combines both neuron types, adopting D-LIF neurons in the 3rd, 4th, 5th, 6th, 7th, and 8th LIF layers, while S-LIF neurons are implemented in the 1st, 2nd, and 9th LIF layers, for the complementary integration of deterministic and stochastic neural dynamics within the single model. The rationale for assigning D-LIF and S-LIF neurons to these specific layers is provided in Section V, where a detailed layer-wise error-tolerance ablation study is presented.

The three Spiking-VGG9 models were implemented considering the hardware-specific characteristics of the proposed CMOS 8T SRAM-based SNN accelerator. All kernel weights were quantized to 3-bit precision to align with the bit resolution of the 8T SRAM array. Additionally, the input pixel values were converted into Poisson spike trains based on the LUT-based spike probability mapping, where the $PC<5:0>$ code corresponding to the closest spike probability to the input pixel value is selected. In the D-LIF neuron layer, a post-synaptic spike is generated only when the membrane potential exceeds the threshold. On the other hand, in the S-LIF neuron layer, the post-synaptic spike is generated probabilistically, following the equation specified in Fig. 4(c).

These VGG-9 models were trained using a surrogate gradient-based backpropagation method [4], with Cross-Entropy loss function for CIFAR-10 classification. ENB<5:0> codes are set as design parameters to control the stochastic behavior of S-LIF neurons depending on the desired trade-off between error tolerance and energy efficiency. For instance, if higher error tolerance is required for a specific layer, a smaller ENB<5:0> value is selected.

Fig. 9(c) presents the error tolerance analysis of the three Spiking-VGG9 models, each operating over 64 time steps. Random weight deviations ranging from 0% to 5% were applied in 0.5% increments, with each test point sampled with 100 test iterations to ensure statistical reliability. The D-SNN model exhibited abrupt accuracy degradation, with the median accuracy degradation to 75.49% at a 5% weight deviation, showing a substantial accuracy drop with 12.59%. In contrast, the S-SNN model suppresses the accuracy degradation with the increasing weight deviation rate, with the median accuracy of 78.42% at a 5% weight deviation point, resulting in a smaller accuracy drop of 9.53%. Notably, the H-SNN model exhibited superior error tolerance, achieving a median accuracy of 80.88% at 5% weight deviation, with an accuracy drop of only 7.67%, outperforming both the D-SNN and S-SNN models.

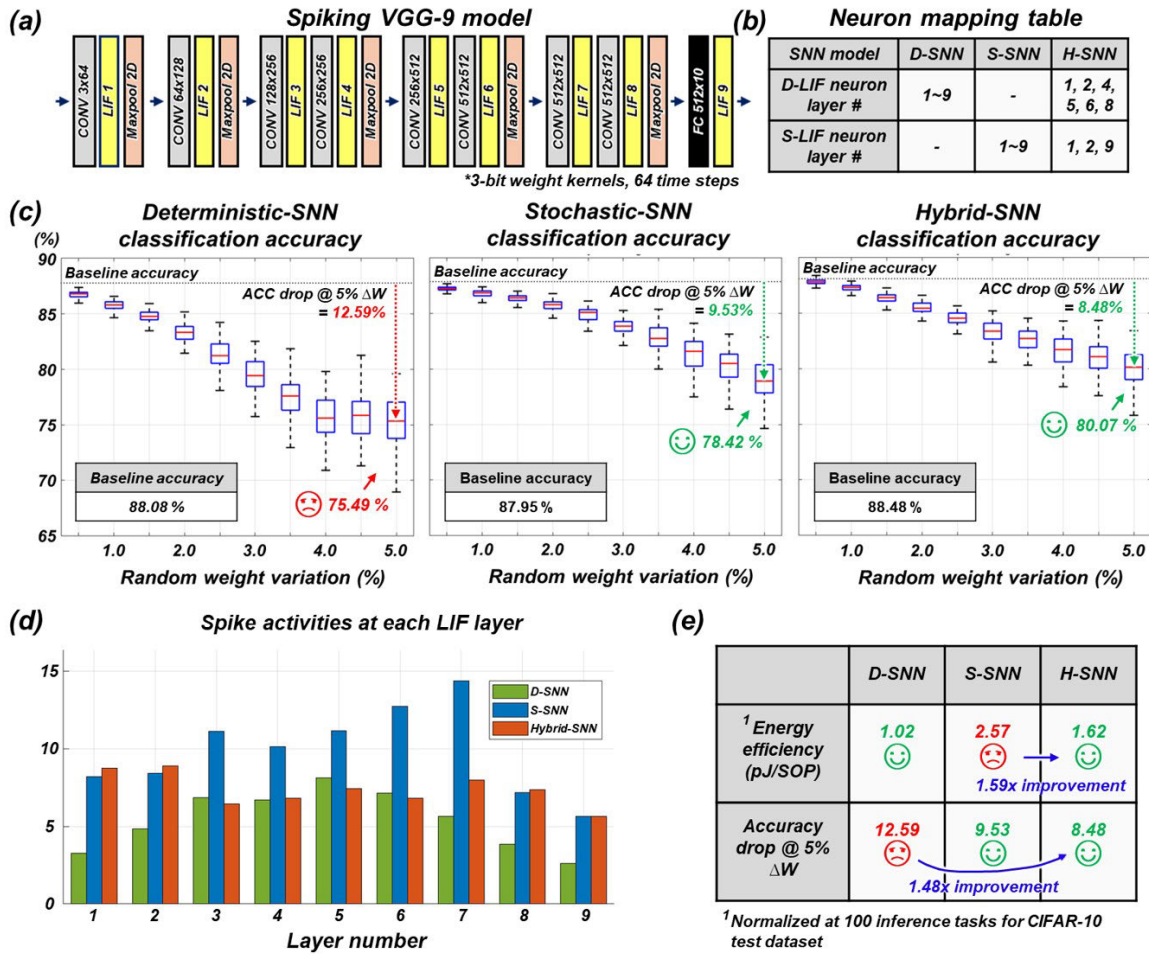


FIGURE 9. (a) Spiking-VGG9 model for CIFAR-10 dataset classification, (b) the LIF neuron type configuration mapping table for three distinct Spiking VGG9 models (D-SNN, S-SNN and H-SNN), (c) Error tolerance evaluation of the three Spiking-VGG9 (D-SNN, S-SNN and H-SNN) models under random weight deviations ranging from 0% to 5%, based on 100 iterations per test point, (d) Layer-wise spike activity analysis of the three Spiking-VGG9 models and (e) Comparative analysis of energy consumption across the three Spiking-VGG9 models.

Fig. 9(d) illustrates the layer-wise spike activity analysis across the three Spiking models over 64 operational time steps. The D-SNN model (green bar) demonstrates the sparsest spike activity across all nine LIF layers, indicating the highest energy efficiency among the three models. In contrast, the S-SNN model (blue bar) shows the most frequent spike activity, corresponding to the lowest energy efficiency. The H-SNN model (orange bar) exhibits slightly higher spike activity than the D-SNN model, but significantly lower spike activity than the S-SNN model.

Fig. 9(e) illustrates the normalized performance metrics of the models in terms of energy efficiency (energy per spike operation) and error tolerance (accuracy drop rate). The energy normalization was performed by correlating the measured power consumption of the fabricated chip and the latency of a single spike operation with the total spike activity for each SNN model. The D-SNN model demonstrates the highest energy efficiency of 1.02 pJ/SOP, but its error tolerance is the lowest, with an accuracy drop of 12.59%.

In contrast, the S-SNN model has the poorest energy efficiency with 2.57 pJ/SOP, yet it achieves a higher error tolerance, with an accuracy drop of 9.53%. While these two models exhibit a clear trade-off between energy efficiency and error tolerance, the proposed H-SNN model achieves an optimal balance, providing improved energy efficiency of 1.62 pJ/SOP, representing a 1.59-fold improvement over the S-SNN model, while maintaining high error tolerance with an accuracy drop of 7.67%, a 1.64-fold improvement over the D-SNN model.

It should be noted that the performance trade-off observed in the proposed Hybrid-SNN primarily originates from the software-level neuron allocation strategy, where the placement of D-LIF and S-LIF neurons determines the balance between energy efficiency and error tolerance. From a hardware perspective, the Hybrid-SNN architecture can support both D-LIF and S-LIF operations within the same hardware platform without requiring additional hardware resources or architectural modifications. This enables flexible

TABLE 1. Comparison with prior works.

	TETC'25 [28]	JSSC'20 [29]	JETCAS'20 [30]	TBCAS'18 [20]	TCASII'21 [21]	TBCAS'23 [8]	TCASI'23 [11]	This work		
Technology [nm]	65	55	65	65	65	55	180	180		
V_{DD} [V]	-	0.8 - 1.0	-	1.3	0.89	1.0	-	1.8		
Computing	Digital NMC	Digital NMC + CIM	Analog CIM	Analog CIM	Digital CIM	Mixed CIM	Mixed CIM	Mixed CIM		
Exp./Sim.	Sim.	Exp.	Sim.	Sim.	Exp.	Exp.	Exp.	Exp.		
Algorithm	DNN	DNN	DNN	SNN	SNN	SNN	SNN	SNN		
Neuron Type	Arithmetic Logic	Accumulator	ADC	Stochastic	D-LIF/S-LIF	D-LIF	D-LIF	D-LIF	S-LIF	H-LIF
Mem. Capacitor [pF]	-	-	-	1	-	-	0.1	4.5 (^c 0.45)		
Area/Neuron [μm^2]	^b 23200	-	12800	700	-	320	-	13260 (^c 7956)		
^a Normalized Area/Neuron [μm^2]	^b 23200	-	12800	700	-	446.94	-	1729 (^c 1037)		
H/W Error tolerance	Δ	Δ	Δ	O	O	Δ	Δ	Δ	O	O
SOP energy [pJ/SOP]	-	-	-	2.14	2.64	5.3	0.53	1.02 (^c 0.65)	2.57 (^c 1.64)	1.62 (^c 1.04)
Energy efficiency [TOPS/W]	0.2	0.6	1.62	0.93	0.76	0.38	3.77	1.96 (^c 3.07)	0.78 (^c 1.22)	1.23 (^c 1.93)
^d Normalized TOPS/W	0.2	0.43	1.62	0.93	0.76	0.27	28.94	15.0 (^c 23.5)	5.97 (^c 9.4)	9.43 (^c 14.8)
Classification Accuracy (%)	-	91.93 (CIFAR-10)	99.13 (MNIST)	94.64 (MNIST)	85.76 (CIFAR-10)	-	-	88.08 (CIFAR-10)	87.95 (CIFAR-10)	88.55 (CIFAR-10)

^a Scaled to 65nm technology node, where squared area normalization is applied per technology node: $A_{norm} = A_{node} (65nm / node)^2$. ^b Accumulator size calculated by evenly dividing the execution-unit area by the number of adders, assuming four 17-bit multipliers and four adders implement one single accumulator logic. ^c Assuming that the membrane capacitor size is optimized by a factor of 1/10 through a proportional reduction in the current mirror size. ^d TOPS/W normalized to a 65 nm technology node, with energy scaled by $(Tech / 65)^2$.

deployment of various Hybrid-SNN configurations while maintaining hardware efficiency.

Finally, the assignment of D-LIF and S-LIF neurons to each layer was empirically determined through trial-and-error to jointly optimize energy efficiency and error tolerance. Too few S-LIF neurons result in reduced error tolerance, while an excessive number of S-LIF neurons leads to an increase in spike activity, thereby degrading energy efficiency. However, this layer-wise assignment procedure can be optimized using a Neural Architecture Search (NAS) framework [31], where hardware constraints such as energy consumption and stochastic error characteristics are jointly considered during optimization, enabling simultaneous optimization of error tolerance and energy efficiency without manual tuning.

D. PERFORMANCE COMPARISON

Table 1 presents a performance comparison with other state-of-the-art DNN and SNN hardware research works. When comparing with SNN hardware works, the majority of other works utilize energy-efficient Deterministic LIF (D-LIF) neurons, which are prone to degradation in classification accuracy under hardware-induced variations, such as PVT variations and random device mismatches.

In contrast, our work introduces a Hybrid LIF (H-LIF) neuron, supporting both (1) Deterministic LIF (D-LIF) mode and (2) Stochastic LIF (S-LIF) mode. This dual-mode configuration not only improves (1) energy efficiency but also enhances (2) error tolerance against hardware-induced variations, particularly when the hybrid mode is strategically distributed across each layer of the SNN model.

Our proposed neuron design exhibits a relatively large area of 13,260 μm^2 , primarily due to two factors: (1) the utilization of a 180 nm technology node, which is the largest among the compared designs, and (2) the use of a relatively

large membrane capacitor with a value of 4.5 pF. When normalizing the neuron area to a 65 nm technology node, the area of the proposed H-LIF neuron decreases to 1,729 μm^2 , although it remains larger than other designs, primarily due to the large membrane capacitor size. Meanwhile, the work in [11] achieves the lowest energy consumption at 0.53 pJ/SOP, mainly due to the minimal size of the membrane capacitor (0.1 pF).

From the perspective of area and energy optimization, there is still a potential for performance improvement of our work by reducing the size of the membrane capacitor. If the size of the mirroring transistors in the MAC current summer (MCS) stage, as shown in Fig. 2(c), is reduced by a factor of 10, the membrane capacitor size can be scaled down by the same factor of 10. With this optimization, the proposed design can achieve a normalized area per neuron of 1,037 μm^2 , along with an improved energy consumption of 1.04 pJ/SOP in Hybrid-SNN mode.

When comparing DNN and SNN hardware in a more comprehensive manner, a clear trade-off can be observed between output activation interface complexity and classification accuracy. SNN hardware typically achieves area-efficient output activation through spiking neuron implementations such as D-LIF, S-LIF, and stochastic neurons, resulting in a normalized area per neuron ranging from 400 to 1,800 μm^2 . In contrast, DNN hardware requires more complex output interfaces, including arithmetic logic units, accumulators, and ADCs, which increase the output inference area overhead to approximately 12,800 to 23,200 μm^2 .

On the other hand, DNN-based hardware generally achieves higher classification accuracy due to their high-precision computation. In contrast, SNN hardware works exhibit relatively lower accuracy, mainly due to immature training methodologies. Specifically, DNN hardware achieves classification accuracies of 91.93% on CIFAR-10

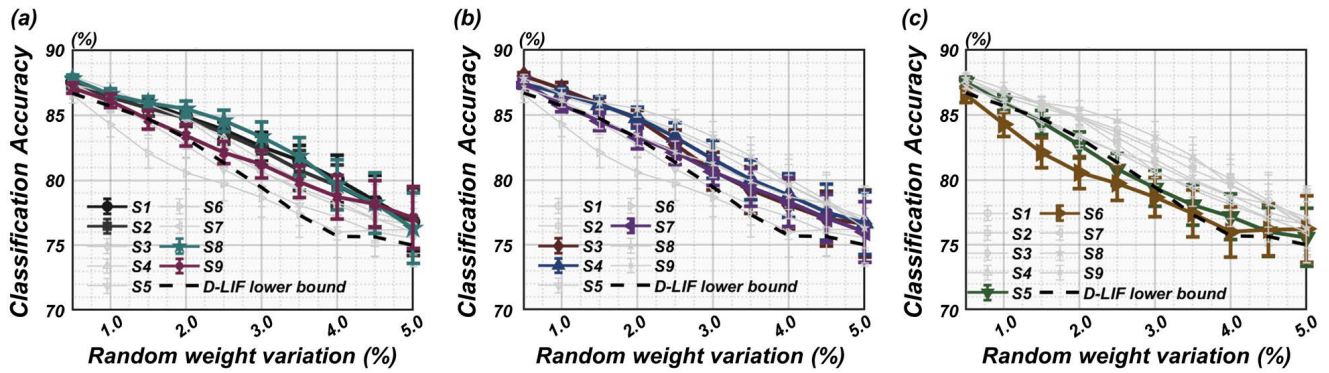


FIGURE 10. Error-tolerance performance analysis of nine Hybrid-SNN configurations, where a single layer of the Spiking-VGG9 network (Layers 1–9) is individually replaced with an S-LIF neuron layer (S1–S9): (a) high-robustness configurations (Layers 1, 2, 8, and 9), (b) moderate-robustness configurations (Layers 3, 4, and 7), and (c) low-robustness configurations (Layers 5 and 6).

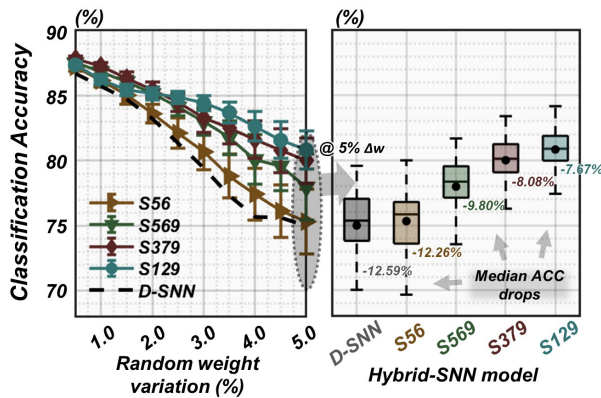


FIGURE 11. Error-tolerance performance analysis of Hybrid-SNN models under four different S-LIF neuron placement configurations.

and 99.13% on MNIST, whereas SNN hardware achieves 85–89% on CIFAR-10 and 94.64% on MNIST.

V. ABLATION STUDY OF HYBRID-SNN MODELS AND STATISTICAL QUALITY ANALYSIS

To further investigate the layer-wise effects of stochasticity injection in Hybrid-SNNs and its impact on error-tolerance performance, we evaluated the Spiking-VGG9 model by replacing each of the nine layers, one at a time, with S-LIF neurons while keeping the remaining layers as D-LIF neurons. We then analyzed the corresponding changes in error-tolerance performance as a function of the specific layer in which stochastic neurons were introduced, as shown in Fig. 10.

When S-LIF neurons are introduced into layers 1, 2, 8, or 9 (S1, S2, S8 and S9), the Hybrid-SNN exhibited the highest robustness against weight variations (Fig. 10(a)). Configurations employing S-LIF neurons in Layers 3, 4, or 7 (S3, S4 and S7) also show improved error tolerance compared with the fully deterministic baseline, but the degree of improvement is relatively moderate (Fig. 10(b)). In contrast, when S-LIF neurons are placed in Layers 5 or 6 (S5 and S6), the error-tolerance performance is the lowest

TABLE 2. 95% Confidence intervals (CIs) and standard deviations of accuracy drop at 5% random weight variation.

Config.	Median Accuracy Drop (%)	95% CI (%)	Std. Dev. (%)
Deterministic-SNN	12.59	[12.02, 13.16]	3.07
Hybrid-SNN-S56	12.26	[11.80, 12.72]	2.47
Hybrid-SNN-S569	9.80	[9.34, 10.26]	2.50
Hybrid-SNN-S379	8.08	[7.75, 8.41]	1.73
Hybrid-SNN-S129	7.67	[7.38, 7.96]	1.46

among all evaluated configurations (Fig. 10(c)), and at certain levels of weight variation even falls below that of the fully deterministic SNN baseline.

Fig. 11 further presents the error-tolerance results for different combinations of S-LIF neuron layer placements, showing different tendencies depending on the selected layers. First, when S-LIF neurons are applied only to Layers 5 and 6—which individually showed very small error-tolerance improvements as shown in Fig. 10(c)—the resulting error-tolerance gain is the smallest among all configurations (median accuracy drop of 12.26% at 5% ΔW). This result remains close to that of the fully deterministic baseline (median accuracy drop of 12.59% at 5% ΔW), providing almost no improvement in error tolerance. When an additional S-LIF neuron is introduced at Layer 9, forming the S569 configuration, the error tolerance slightly increased (median accuracy drop of 9.80% at 5% ΔW). The S379 configuration, demonstrates substantially better performance (median accuracy drop of 8.08% at 5% ΔW). Finally, the S129 configuration achieves the highest error tolerance among all evaluated cases (median accuracy drop of 7.67% at 5% ΔW). Overall, these results indicate that the effectiveness of stochasticity injection for improving error tolerance varies depending on the layer in which S-LIF neurons are introduced.

To further evaluate the statistical significance of the observed error-tolerance improvement, 95% confidence intervals (CIs) of the accuracy drop were calculated using the Student's t-distribution based on 100 independent iterations

for the Deterministic-SNN (D-SNN) baseline and the four Hybrid-SNN configurations. The Confidence Interval (CI) is defined as

$$CI = \bar{x} \pm t(0.975, N - 1) \frac{s}{\sqrt{N}} \quad (1)$$

where $\bar{x}$ denotes the mean accuracy drop, $t(0.975, N - 1)$ is the critical value of the Student's t-distribution at the 95% confidence level, s is the sample standard deviation, and N is the number of iterations.

Table 2 summarizes the CIs and standard deviations under the 5% random weight-variation condition of the D-SNN and four different Hybrid-SNN models. The Hybrid-SNN-S569, S379, and S129 configurations exhibit confidence intervals that do not overlap with that of the D-SNN baseline, suggesting a statistically distinguishable reduction in accuracy degradation under random weight variations. In contrast, the Hybrid-SNN-S56 configuration exhibits substantial overlap with the baseline confidence interval, indicating that its improvement in error tolerance is relatively limited.

The variability of the accuracy drop also differs across the evaluated configurations. The D-SNN baseline model exhibits the largest standard deviation of 3.07%, whereas the Hybrid-SNN-S56 and S569 configurations reduce the variation to 2.47% and 2.50%, respectively. More reductions in variance are observed for the Hybrid-SNN-S379 and S129 configurations, which achieve standard deviations of 1.73% and 1.46%, respectively. These results indicate that the reduction in performance variability under random weight perturbations is also dependent on the selected layer for S-LIF neuron placement.

VI. CONCLUSION

This work introduces a CMOS-based SNN accelerator with the Hybrid Leaky Integrate-and-Fire (H-LIF) neuron, capable of operating in both (1) Deterministic LIF (D-LIF) and (2) Stochastic LIF (S-LIF) modes. This dual-mode functionality addresses a critical gap in current hardware by enabling the concurrent support of (1) energy-efficient and (2) error-tolerant neuron operations on the same hardware platform. By strategically integrating both neuron types into the Spiking VGG-9 model, the Hybrid-SNN achieves enhanced energy efficiency, consuming 1.62 pJ per spike operation (SOP), which results in a 1.59x improvement in energy efficiency compared to the Stochastic-SNN (S-SNN). Additionally, the Hybrid-SNN exhibits a reduced accuracy degradation of 7.67%, providing a 1.64x improvement in robustness compared to the Deterministic-SNN (D-SNN). This approach effectively exploits the advantages of both D-LIF and S-LIF neurons, representing a significant step forward in the development of versatile and resilient SNN hardware.

REFERENCES

- [1] E. M. Izhikevich, "Simple model of spiking neurons," *IEEE Trans. Neural Netw.*, vol. 14, no. 6, pp. 1569–1572, Nov. 2003.
- [2] M. Pfeiffer and T. Pfeil, "Deep learning with spiking neurons: Opportunities and challenges," *Frontiers Neurosci.*, vol. 12, p. 18, Oct. 2018. [Online]. Available: <https://www.frontiersin.org/articles/10.3389/fnins.2018.00774>
- [3] K. Roy, A. Jaiswal, and P. Panda, "Towards spike-based machine intelligence with neuromorphic computing," *Nature*, vol. 575, no. 7784, pp. 607–617, Nov. 2019, doi: [10.1038/s41586-019-1677-2](https://doi.org/10.1038/s41586-019-1677-2).
- [4] J. K. Eshraghian, M. Ward, E. O. Neftci, X. Wang, G. Lenz, G. Dwivedi, M. Bennamoun, D. S. Jeong, and W. D. Lu, "Training spiking neural networks using lessons from deep learning," *Proc. IEEE*, vol. 111, no. 9, pp. 1016–1054, Sep. 2023.
- [5] F. Akopyan, J. Sawada, A. Cassidy, R. Alvarez-Icaza, J. Arthur, P. Merolla, N. Imam, Y. Nakamura, P. Datta, G.-J. Nam, B. Taba, M. Beakes, B. Brezzo, J. B. Kuang, R. Manohar, W. P. Risk, B. Jackson, and D. S. Modha, "TrueNorth: Design and tool flow of a 65 mW 1 million neuron programmable neurosynaptic chip," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 34, no. 10, pp. 1537–1557, Oct. 2015.
- [6] M. Davies et al., "Loihi: A neuromorphic manycore processor with on-chip learning," *IEEE Micro*, vol. 38, no. 1, pp. 82–99, Jan. 2018.
- [7] J. Kim, J. N. Kim, Y. Kim, S. Hwang, and M. Koo, "Design of a 180 nm CMOS neuron circuit with soft-reset and underflow allowing for loss-less hardware spiking neural networks," *Adv. Intell. Syst.*, vol. 6, no. 1, 2024, Art. no. 2300460. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/aisy.202300460>
- [8] X. Qi, J. Zhao, Y. Lou, G. Wang, K.-T. Tang, and Y. Li, "A 5.3 pJ/Spike CMOS neural array employing time-modulated axon-sharing and background mismatch calibration techniques," *IEEE Trans. Biomed. Circuits Syst.*, vol. 17, no. 2, pp. 286–298, Apr. 2023.
- [9] A. Valentian, F. Rummens, E. Vianello, T. Mesquida, C. L. de Boissac, O. Bichler, and C. Reita, "Fully integrated spiking neural network with analog neurons and RRAM synapses," in *IEDM Tech. Dig.*, Dec. 2019, pp. 14.3.1–14.3.4.
- [10] S. K. Das, S. K. Nandi, C. V. Marquez, A. Rúa, M. Uenuma, S. K. Nath, S. Zhang, C.-H. Lin, D. Chu, T. Ratcliff, and R. G. Elliman, "Dynamics of leaky integrate-and-fire neurons based on oxyvanite memristors for spiking neural networks," *Adv. Intell. Syst.*, vol. 6, no. 11, 2024, Art. no. 2400191. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/aisy.202400191>
- [11] Y. Liu, Y. Ma, W. He, Z. Wang, L. Shen, J. Ru, R. Huang, and L. Ye, "An 82-nW 0.53-pJ/SOP clock-free spiking neural network with 40-μs latency for AIoT wake-up functions using a multilevel-event-driven bionic architecture and computing-in-memory technique," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 70, no. 8, pp. 3075–3088, Aug. 2023.
- [12] A. Faisal, L. Selen, and D. Wolpert, "Noise in the nervous system," *Nature Rev. Neurosci.*, vol. 9, pp. 292–303, May 2008.
- [13] W. Maass, "Noise as a resource for computation and learning in networks of spiking neurons," *Proc. IEEE*, vol. 102, no. 5, pp. 860–880, May 2014.
- [14] 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 Comput. Biol.*, vol. 7, no. 11, Nov. 2011, Art. no. e1002211, doi: [10.1371/journal.pcbi.1002211](https://doi.org/10.1371/journal.pcbi.1002211).
- [15] C. Gulcehre, M. Moczulski, M. Denil, and Y. Bengio, "Noisy activation functions," 2016, *arXiv:1603.00391*.
- [16] E. O. Neftci, B. U. Pedroni, S. Joshi, M. Al-Shedivat, and G. Cauwenberghs, "Stochastic synapses enable efficient brain-inspired learning machines," *Frontiers Neurosci.*, vol. 10, p. 16, Jun. 2016.
- [17] Y. Jiang, S. Lu, and A. Sengupta, "Stochastic spiking neural networks with first-to-spike coding," 2024, *arXiv:2404.17719*.
- [18] I. Yeo, M. Chu, and B.-G. Lee, "A power and area efficient CMOS stochastic neuron for neural networks employing resistive crossbar array," *IEEE Trans. Biomed. Circuits Syst.*, vol. 13, no. 6, pp. 1678–1689, Dec. 2019.
- [19] T. Tuma, A. Pantazi, M. Le Gallo, A. Sebastian, and E. Eleftheriou, "Stochastic phase-change neurons," *Nature Nanotechnol.*, vol. 11, no. 8, pp. 693–699, Aug. 2016.
- [20] J. Lin and J.-S. Yuan, "Analysis and simulation of capacitor-less ReRAM-based stochastic neurons for the in-memory spiking neural network," *IEEE Trans. Biomed. Circuits Syst.*, vol. 12, no. 5, pp. 1004–1017, Oct. 2018.
- [21] Y. Kuang, X. Cui, Y. Zhong, K. Liu, C. Zou, Z. Dai, Y. Wang, D. Yu, and R. Huang, "A 64K-neuron 64M-1b-synapse 2.64 pJ/SOP neuromorphic chip with all memory on chip for spike-based models in 65 nm CMOS," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 68, no. 7, pp. 2655–2659, Jul. 2021.

- [22] H. Jeon, J. Kim, Y. Kim, and M. Koo, "ES-SNN: E-flash memory-based stochastic spiking neural networks for edge inference," *IEEE Access*, vol. 14, pp. 35894–35900, 2026.
- [23] H. Das, K. P. Patel, R. D. Febbo, C. D. Schuman, and G. S. Rose, "Leveraging stochasticity in memristive synapses for efficient and reliable neuromorphic systems," *Npj Unconventional Comput.*, vol. 2, no. 1, p. 3, Feb. 2025, doi: [10.1038/s44335-024-00017-x](https://doi.org/10.1038/s44335-024-00017-x).
- [24] M. Koo, G. Srinivasan, Y. Shim, and K. Roy, "SBSNN: Stochastic-bits enabled binary spiking neural network with on-chip learning for energy efficient neuromorphic computing at the edge," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 67, no. 8, pp. 2546–2555, Aug. 2020.
- [25] M.-F. Chang, C.-F. Chen, T.-H. Chang, C.-C. Shuai, Y.-Y. Wang, and H. Yamauchi, "17.3 A 28 nm 256 kb 6T-SRAM with 280 mV improvement in VMIN using a dual-split-control assist scheme," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2015, pp. 1–3.
- [26] H. Kim, Y. An, M. Kim, G.-C. Heo, and Y. Shim, "All stochastic-spiking neural network (AS-SNN): Noise induced spike pulse generator for input and output neurons with resistive synaptic array," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 72, no. 1, pp. 78–82, Jan. 2025.
- [27] T. Rabuske and J. Fernandes, *Charge-Sharing SAR ADCs for Low-Voltage Low-Power Applications*. Cham, Switzerland: Springer, 2018. [Online]. Available: <https://books.google.co.kr/books?id=VSGeuAEACAAJ>
- [28] M. Caon, C. Choné, P. D. Schiavone, A. Levisse, G. Masera, M. Martina, and D. Atienza, "Scalable and RISC-V programmable near-memory computing architectures for edge nodes," *IEEE Trans. Emerg. Topics Comput.*, vol. 13, no. 3, pp. 1003–1018, Jul. 2025.
- [29] Y.-C. Chiu, Z. Zhang, J.-J. Chen, X. Si, R. Liu, Y.-N. Tu, J.-W. Su, W.-H. Huang, J.-H. Wang, W.-C. Wei, J.-M. Hung, S.-S. Sheu, S.-H. Li, C.-I. Wu, R.-S. Liu, C.-C. Hsieh, K.-T. Tang, and M.-F. Chang, "A 4-kb 1-to-8-bit configurable 6T SRAM-based computation-in-memory unit-macro for CNN-based AI edge processors," *IEEE J. Solid-State Circuits*, vol. 55, no. 10, pp. 2790–2801, Oct. 2020.
- [30] A. Agrawal, A. Kosta, S. Kodge, D. E. Kim, and K. Roy, "CASH-RAM: Enabling in-memory computations for edge inference using charge accumulation and sharing in standard 8T-SRAM arrays," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 10, no. 3, pp. 295–305, Sep. 2020.
- [31] R. Vidya Wicaksana Putra and M. Shafique, "NeuroNAS: Enhancing efficiency of neuromorphic in-memory computing for intelligent mobile agents through hardware-aware spiking neural architecture search," 2024, *arXiv:2407.00641*.



HONGGU KIM received the B.S. degree from the School of Electrical and Electronics Engineering, Chung-Ang University (CAU), Seoul, South Korea, in 2022. He is currently pursuing the integrated M.S. and Ph.D. degree in intelligent semiconductor engineering. His research interests include neuromorphic hardware and software co-optimization and analog compute-in-memory architecture.



YERIM AN received the B.S. degree from the School of Electrical and Electronics Engineering, Tech University of Korea (TUKOREA), South Korea, in 2022, and the M.S. degree in intelligent semiconductor engineering from Chung-Ang University (CAU), Seoul, South Korea, in 2024, where she is currently pursuing the Ph.D. degree in intelligent semiconductor engineering. Her research interest includes process-in-memory (PIM) architecture design.



JAEOYOUN KIM received the B.S. degree from the School of Electrical and Electronics Engineering, Chung-Ang University (CAU), Seoul, South Korea, in 2024. He is currently pursuing the integrated M.S. degree in intelligent semiconductor engineering. His research interests include DRAM-based analog compute-in-memory architecture and CMOS interface design for monolithic 3D DRAM-CMOS integrated systems.



YONG SHIM (Member, IEEE) received the B.S. and M.S. degrees in electronics engineering from Korea University, in 2004 and 2006, respectively, and the Ph.D. degree from the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, in 2018. He was a Memory Interface Designer with Samsung Electronics, Hwaseong, from 2006 to 2013. At Samsung, he has worked on the design and development of a memory interface for synchronous DRAMs (DDR1 and DDR4). He is currently an Assistant Professor with Chung-Ang University. Prior to joining Chung-Ang University, in 2020, he was an SRAM Designer with Intel Corporation, Hillsboro, OR, from 2018 to 2020, where he was involved in designing circuits for super-scaled next-generation SRAM cache design. His research interests include neuromorphic hardware and algorithm, in-memory computing, robust memory interface design, and emerging devices (RRAM, MRAM, and STO)-based unconventional computing models.

...