Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Hardware-Aware On-Chip Learning with a Scalable Single-Transistor-based Neuromorphic Platform

Sangbum Kim

sangbum.kim@snu.ac.kr

Seoul National University https://orcid.org/0000-0001-7460-3750

Sung Min Lee

Seoul Natational University

Jae-Hyun Lee

Seoul National University

Weonjun Heo

Seoul National University

Seonjae Park

Seoul National University

Min-Hyeok Jo

Seoul National University

Jun-Young Park

Seoul National University https://orcid.org/0009-0006-5439-0384

Seung-heon Chris Baek

Korea Institute of Science and Technology https://orcid.org/0000-0003-1338-8238

Joon-Kyu Han

Seoul National University https://orcid.org/0000-0002-8736-9091

Taelim Heo

Seoul National University

Article

Keywords:

Posted Date: November 5th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11537086/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

Hardware-Aware On-Chip Learning with a Scalable Single-Transistor-based Neuromorphic Platform

Sungmin Lee[1,2,3†], Jae-Hyun Lee[1,2†], Weonjun Heo[1,2,3], Taelim Heo[1,2,3], Seonjae Park[1,2], Min-Hyeok Jo[1,2], Jun-Young Park[1,2], Seung-heon Chris Baek[4], Joon-Kyu Han[1,2*], Sangbum Kim[1,2,3*]

[1*]Department of Materials Science and Engineering, College of Engineering, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul, 08826, Republic of Korea.

[2]Inter-University Semiconductor Research Center, Seoul National University, 104-dong, 1 Gwanak-ro, Gwanak-gu, Seoul, 08826, Republic of Korea.

[3]Research Institute of Advanced Materials, Seoul National University, Bldg. 131, 1 Gwanak-ro, Gwanak-gu, Seoul, 08826, Republic of Korea.

[4]Center for Semiconductor Technology, Korea Institute of Science and Technology, Seoul, 02792, Republic of Korea.

*Corresponding author(s). E-mail(s): joonkyuhan@snu.ac.kr; sangbum.kim@snu.ac.kr;

†These authors contributed equally to this work.

Abstract

Spiking neural networks (SNNs) offer a promising paradigm for energy-efficient computing, yet enabling on-chip training in analog hardware platforms remains challenging because internal neuronal states are difficult to access. Here, we present a hardware–algorithm co-designed neuromorphic architecture for state-access-free learning using the accessible signals of a single-transistor (1T) platform. Current-based surrogate gradient (cSG) and eligible leaky integrate-and-fire encoder (eLIFE) modules generate learning signals from spike events without explicit membrane-potential readout. Co-integrated 1T-synapses, 1T-neurons, and 1T-cSG devices demonstrate synaptic weight storage, leaky integrate-and-fire neuronal dynamics, and surrogate gradient signal generation in a measured unit architecture. At the algorithmic level, simulations demonstrate competitive performance using small-scale networks on F-MNIST, S-MNIST, and SHD and large-scale networks on CIFAR-10, CIFAR-100, CIFAR10DVS, DVS128Gesture, ImageNet, and COCO. Furthermore, hardware-aware simulations of an analog on-chip continual-learning framework demonstrate learning of previously unseen classes with limited average accuracy loss on previously learned classes. Under the adopted model, estimated inference energy is more than 60% lower than that of fully digital SNN baselines. These results provide a unified device–algorithm framework for state-access-free learning in adaptive neuromorphic systems.

Neuromorphic computing has emerged as a scalable paradigm for addressing the escalating computational and energy demands of modern AI systems[1]. At the algorithmic core of neuromorphic computing, spiking neural networks (SNNs) emulate biological communication by encoding information as asynchronous binary spikes, exploiting event-driven parallelism for ultra-low-power operation[2]. In parallel, neuromorphic engineering has produced a broad spectrum of hardware platforms, spanning purely digital processors[3–8], mixed-signal chips[9,10], and advanced analog computing substrates[11,12]. Such hardware diversity provides the physical foundation for spike-based computation ranging from resource-constrained edge applications to high-performance computing servers.

The binary nature of SNNs has motivated their integration into digital in-memory computing (DIMC) architectures. A key advantage of spiking DIMC lies in replacing power-intensive multiply-accumulate (MAC) operations with simple accumulate (AC) operations[13]. This architectural simplification offers a practical route toward scalable neuromorphic systems and has driven extensive research on dataflow optimization and quantization techniques for improving power efficiency[14]. Nevertheless, purely digital implementations remain constrained by static and dynamic power dissipation. Even with aggressive architectural optimization, digital SNN processors often exhibit a practical power floor at the microwatt (μW) scale[15], motivating a transition toward analog implementations for further energy reduction.

In analog SNN implementations, analog in-memory computing (AIMC) provides an energy-efficient means of realizing synaptic computation by performing vector–matrix multiplication (VMM) directly in the analog domain, thereby reducing computational and data-movement overhead. The resulting continuous analog VMM outputs must then be converted into spike-domain signals for subsequent event-driven processing. Spiking neurons serve this role by integrating the analog synaptic outputs and generating asynchronous, time-encoded spikes, thereby enabling efficient communication between successive SNN layers while preserving the energy advantage of AIMC[16].

Among various spiking neuron models, the leaky integrate-and-fire (LIF) model is widely used because of its simple dynamics and analog compatibility. Conventional circuit-based LIF neurons, however, typically require multiple transistors, resulting in considerable area and power overhead[17–19]. To overcome these limitations, single-device neurons that exhibit abrupt switching behavior have attracted significant attention[20–23]. For large-scale implementation, however, compact neuronal functionality alone is insufficient; the underlying device technology should also be compatible with established semiconductor fabrication and peripheral circuitry. From this perspective, floating-body (FB) MOSFET-based single-transistor (1T) neurons offer a promising approach by exploiting floating-body charge accumulation and regenerative switching to realize intrinsic integrate-and-fire behavior within a mature CMOS-compatible platform[24–26]

Moreover, the FB MOSFET platform can be extended to enable homogeneous integration of neurons and synapses. Incorporating a charge-trap layer (CTL) provides nonvolatile post-fabrication control over neuronal firing characteristics, while the same charge-storage mechanism can be utilized to realize flash-memory-based 1T-synapses. This functional versatility enables wafer-scale co-integration of compact neurons and

synaptic devices within a common transistor platform. However, existing demonstrations of such co-integrated systems have remained largely limited to oscillatory modes, which are not directly compatible with VMM-based AIMC operation, leaving a critical gap toward practical analog SNN systems[25,26].

Beyond hardware implementation, effective training strategies are essential for optimizing SNNs and compensating for device- and circuit-level nonidealities in analog systems[12]. Among diverse training approaches, surrogate-gradient-based direct training has emerged as a dominant paradigm, enabling high performance on complex tasks[27]. However, scaling SNN training modules remains challenging because recurrent neuronal dynamics give rise to a spatio-temporal credit-assignment problem. Backpropagation through time (BPTT) addresses the spatio-temporal credit-assignment problem by unrolling neuronal dynamics over time[28], but requires the storage of intermediate states and consequently incurs substantial memory and computational overhead. Eligibility-trace-based methods reduce such overhead by localizing temporal credit assignment[29,30], while spatial learning through time (SLTT)[31] and spatial online learning at once (SOLO)[32] rely on spatial gradients computed at every time step and only at the final time step, respectively (See details in Supplementary Note S1.1). Nevertheless, hardware implementation of on-chip learning remains constrained by the need to access internal neuronal states during gradient computation[33]. Extracting the internal states requires additional sensing and readout circuitry, introducing substantial area, energy, and communication overhead. Accordingly, translating efficient learning rules into practical hardware requires hardware–algorithm co-design that considers not only computational complexity but also the interdependence of computational primitives, the accessibility of required variables, and the associated dataflow and communication overheads[34,35].

In this work, we present a hardware–algorithm co-designed analog SNN architecture incorporating state-access-free on-chip learning modules. At the algorithm level, the proposed current-based surrogate gradient (cSG) and eligible leaky integrate-and-fire encoding (eLIFE) provide hardware-friendly alternatives to learning signals that otherwise require membrane-potential readout. The cSG derives surrogate signals from locally available synaptic currents, whereas eLIFE generates spike-based eligibility signals through auxiliary recurrent neuronal dynamics. Benchmark comparisons with conventional learning methods demonstrate competitive accuracy and training capability. At the hardware level, 1T-synapses, 1T-neurons, and 1T-cSG modules are homogeneously integrated on a wafer-scale platform, demonstrating synapse-driven LIF dynamics and device-level surrogate-gradient generation. At the system level, simulations incorporating measured device characteristics validate scalable analog-tile operation and demonstrate on-chip fine-tuning across static and temporal tasks. The proposed architecture maintains competitive accuracy without explicit access to internal analog neuronal states, providing a common-device basis for hardware-compatible SNN learning. Under the adopted model, inference energy is estimated to be more than 60% lower than for fully digital SNN implementations.

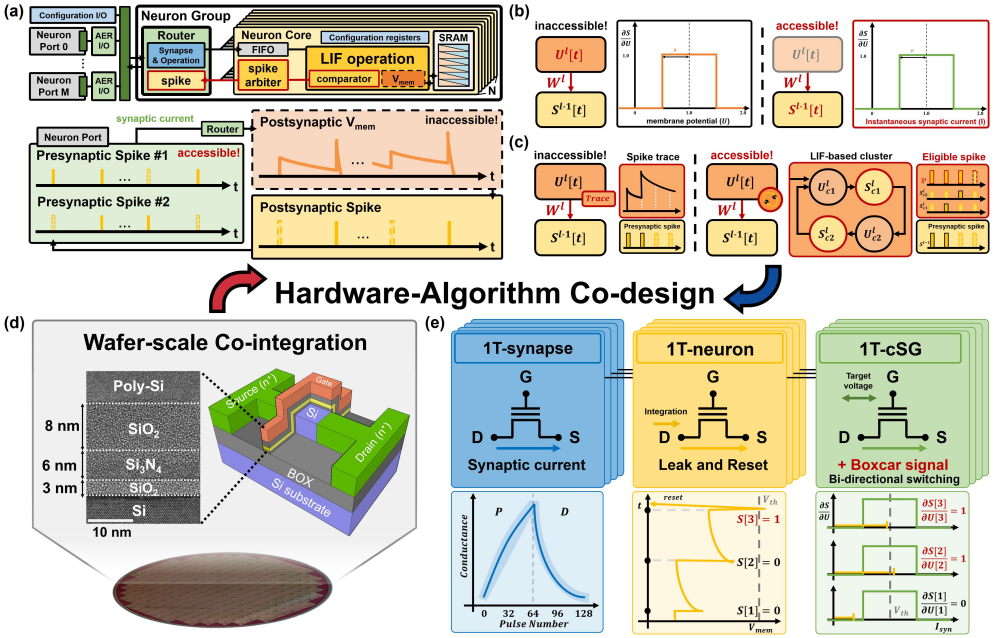


Fig. 1 Overview of the neuromorphic computing system and state-access problem in training SNNs. (a) System block diagram of neuromorphic computing chips. Within the neuron core, the membrane potential is physically inaccessible; communication is strictly restricted to spike signals via routers and address I/Os. (b) Comparison between the conventional membrane-potential-based surrogate gradient and the proposed current-based surrogate gradient (cSG). (c) Comparison between a conventional eligibility trace, which requires external access and storage of continuous neuronal states, and the proposed eligible leaky integrate-and-fire encoder (eLIFE) module which encapsulates the temporal state evolution within an interconnected LIF-based clone-neuron cluster and exposes only discrete eligible spikes. (d) Schematic illustration and cross-sectional transmission electron microscopy (TEM) image of the homogeneous single-transistor (1T) devices fabricated through wafer-scale co-integration on an 8-inch SOI wafer. (e) Functional building blocks of the co-integrated architecture: a 1T-synapse providing multilevel synaptic conductance through potentiation and depression, a floating-body 1T-neuron performing leaky integration, firing, and reset, and a bidirectionally switching 1T-cSG generating a boxcar surrogate signal according to the instantaneous synaptic current.

Results

Hardware-algorithm co-design with state-access-free modules

As shown in Fig. 1a, most neuromorphic chips typically provide access to synaptic weights and spike events, whereas membrane potentials remain internal neuronal states that are difficult to read during operation (See details in Supplementary Note S1.2). Direct membrane-potential readout would require dedicated circuitry, additional memory, high-bandwidth data paths, continuous sampling, or per-neuron analog-to-digital converters[10]. Consequently, the dependence of conventional surrogate-gradient methods on membrane-potential variables creates a mismatch between algorithmic requirements and hardware-accessible information (Fig. 1b, left). Conventional eligibility-trace implementations reduce the computational cost of temporal credit

assignment but still rely on auxiliary potential states for spike-trace generation (Fig. 1c, left). Because such state-access requirements pose a practical challenge for scalable on-chip training, learning modules must derive weight updates from locally available synaptic weights and spike events without explicit readout of membrane potentials or auxiliary trace-related states[36].

To overcome the state-access bottleneck, two complementary state-access-free learning modules are introduced: a current-based surrogate gradient (cSG) and an eligible leaky integrate-and-fire encoder (eLIFE). Unlike conventional surrogate-gradient methods, which derive surrogate signals from neuronal membrane potentials, the proposed cSG generates surrogate signals from the instantaneous synaptic current produced by weighted presynaptic spike activity (Fig. 1b, right). Because the readout of synaptic current is locally available at the output of the synaptic array, cSG avoids explicit membrane-potential readout and the associated data-conversion and storage circuitry. In parallel, eLIFE addresses the state-access requirement associated with eligibility traces. The eLIFE module embeds temporal dynamics within an interconnected cluster of LIF neurons and converts the resulting eligibility information into discrete eligible spikes (Fig. 1c, right; see Supplementary Note S1.3). Consequently, only spike events are exposed to the external learning path, while the continuous internal states of the LIF-based cluster remain locally encapsulated.

To physically realize the proposed state-access-free framework, 1T-synapses, 1T-neurons, and 1T-cSG devices were co-integrated on an 8-inch silicon-on-insulator (SOI) wafer using a process-compatible 1T-device platform (Fig. 1d; see Supplementary Note S2.1 and Methods "Device fabrication"). Within this platform, each device is configured to perform a distinct function. The nitride CTL incorporated in silicon-oxide-nitride-oxide-silicon (SONOS) gate stack stores multilevel conductance states in the 1T-synapse. The 1T-neuron exploits floating-body charge dynamics and single-transistor latch (STL) phenomenon to implement leaky integration, threshold firing, and reset, while the CTL provides nonvolatile calibration of its firing characteristics. In addition, 1T-cSG exploits gate-controlled bidirectional threshold switching to generate the boxcar surrogate signal from the synaptic current. Thus, synaptic storage, neuronal dynamics, and surrogate-signal generation are integrated on a common, fabrication-compatible platform despite their distinct functionalities (Fig. 1e).

This hardware–algorithm co-design separates externally accessible learning signals—synaptic currents and spike events—from inaccessible internal neuronal states. Consequently, surrogate-gradient and eligibility computations can be performed without explicit membrane-potential readout, providing a scalable foundation for state-access-free on-chip learning.

Device characteristics of 1T-devices

The homogeneous 1T-device platform comprises a 1T-synapse, a 1T-neuron, and a 1T-cSG. Among these components, the 1T-synapse stores synaptic weights through charge trapping in the CTL. The trapped charge modulates the threshold voltage (V_{th}) of the transistor, allowing the channel conductance to represent the synaptic weight, as illustrated in Fig. 2a. As shown in Fig. 2b, electron trapping in the CTL progressively

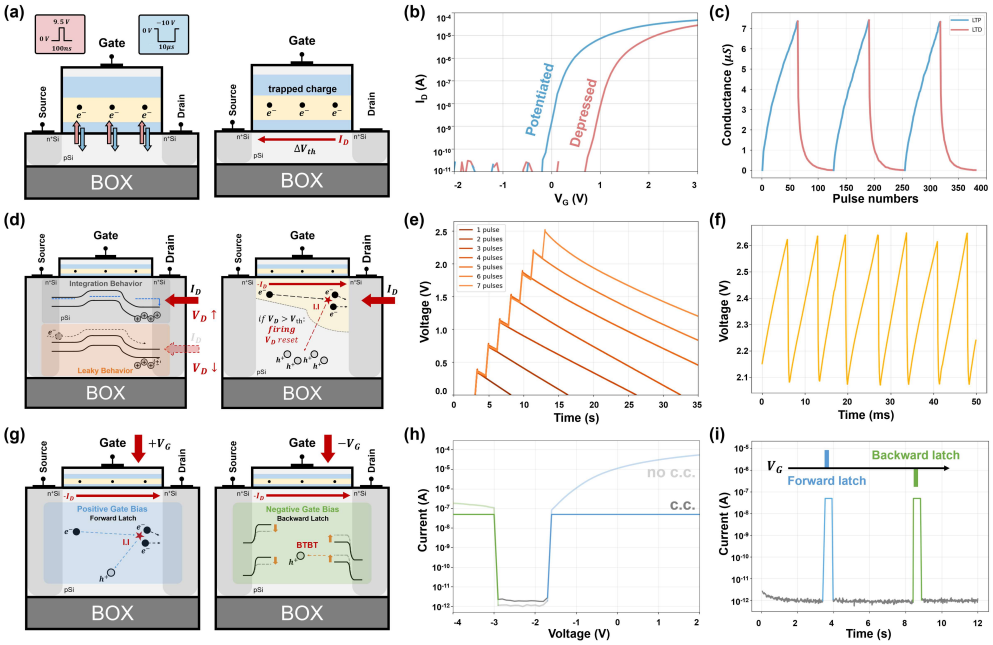


Fig. 2 Device characteristics of the homogeneous 1T-synapse, 1T-neuron, and 1T-cSG platform. (a) Schematic illustration of charge trap-based weight modulation in the 1T-synapse. Gate pulses modulate the trapped charge in the CTL, resulting in a V_{th} shift and corresponding modulation of the channel conductance. (b) Measured transfer characteristics of the 1T-synapse, showing the V_{th} shift induced by charge trapping in the CTL. (c) Potentiation and depression characteristics of the 1T-synapse, demonstrating 64 programmable conductance states over three consecutive cycles. (d) Schematic illustration of the 1T-neuron operation. Gradual charge accumulation and leakage enable leaky integration, whereas impact-ionization-induced single-transistor latch (STL) operation produces abrupt firing and reset. (e) Measured membrane voltage (V_{mem}) transients under one to seven input-current pulses, showing stepwise integration during current injection and gradual leakage during the pulse-off intervals. (f) Repetitive firing characteristics of the 1T-neuron under continuous current injection, showing integration of V_{mem} followed by abrupt reset at each firing event induced by STL phenomenon. (g) Schematic illustration of gate-controlled bidirectional STL operation for the 1T-cSG. A positive-going gate pulse induces forward latch through impact ionization, whereas a negative-going gate pulse induces backward latch through band-to-band tunneling (BTBT). (h) Measured I_{D}-V_{G} characteristics at a V_{D} of 4 V with and without current compliance, showing a low-current region bounded by the backward and forward latch transitions and a compliance-limited output after latching. (i) Transient I_{D} response under positive- and negative-going gate-voltage pulses at a drain voltage of 4 V, demonstrating selective forward and backward latch events, respectively.

shifts V_{th} toward higher gate voltage (V_{G}), resulting in a corresponding modulation of the channel current. By controlling the amount of trapped charge, multiple distinguishable conductance states can therefore be obtained. Fig. 2c demonstrates 64 programmable conductance states over three consecutive potentiation–depression cycles, confirming the multilevel weight-storage capability and the repeatability of the synaptic update behavior. These weight states were programmed using gate pulses of 10 V for 10 μs and +9.5 V for 100 ns for potentiation and depression, respectively.

To further assess device-level robustness, the potentiation and depression characteristics were evaluated across eight devices, demonstrating reproducible weight-update behavior and device-to-device uniformity (See details in Supplementary Note S2.2).

For the neuron operation, V_{G} was fixed at 3 V to maintain the channel in a nominally off state. When an input current (I_{in}) pulse is applied to the drain of the 1T-neuron, positive charge accumulates at the drain electrode due to the parasitic capacitance (C_{par}), thereby increasing the drain voltage (V_{D}), which corresponds to the membrane potential (V_{mem}) of the 1T-neuron. When I_{in} is removed, electrons injected from the source through the off-state leakage path compensate the stored positive charge and gradually discharge the drain electrode, producing leaky decay of the membrane potential (Fig. 2d, left). Consistent with the charge-based mechanism, the measured traces show a stepwise increase in membrane potential during each input pulse and partial decay between successive pulses (Fig. 2e). Increasing the number of input pulses from one to seven progressively raises the membrane potential while preserving inter-pulse leakage. Measurements across eight devices further confirm the device-level reproducibility of leaky integration (See details in Supplementary Note S2.3).

Moreover, the 1T-neuron implements threshold firing and reset through the STL phenomenon[37]. A constant 20 nA I_{in} raises the drain voltage until it reaches the latch-up voltage (V_{latch}), which corresponds to the firing voltage (V_{firing}) of the neuron. The high electric field near the drain induces impact ionization (I.I), and the generated holes trigger regenerative positive feedback, abruptly switching the device to a high-current state (Fig. 2d, right). The threshold-switching behavior is evident in the measured I_{D}-V_{D} characteristics, which exhibit an abrupt increase in I_{D} at V_{latch} (See details in Supplementary Note S2.4) The resulting high current increase rapidly discharges C_{par}, thereby resetting V_{mem}. As shown in Fig. 2f, continuous current injection therefore produces repeated gradual increases in V_{mem} followed by abrupt voltage drops, demonstrating repetitive integration, firing, and reset operation of the 1T-neuron.

Importantly, charge stored in the CTL electrostatically modulates V_{latch} and, consequently, V_{firing}. Electron trapping shifts V_{firing} toward higher voltages, whereas hole trapping shifts it toward lower voltages. This nonvolatile modulation enables post-fabrication calibration of neuronal firing characteristics and mitigates device-to-device variation (See details in Supplementary Note S2.5).

The STL phenomenon also implements the cSG function through gate-controlled bidirectional switching. Under a high V_{D}, increasing V_{G} enhances electron injection from the source and induces I.I near the drain. The generated holes accumulate in the floating body and trigger the forward latch (Fig. 2g, left). In contrast, decreasing the gate voltage to a more negative level strengthens the electric field at the body–drain junction and promotes band-to-band tunneling (BTBT). Hole generation through BTBT triggers the backward latch (Fig. 2g, right). The forward and backward latch transitions are therefore initiated through different carrier-generation mechanisms but share the same regenerative positive feedback.

As shown in Fig. 2h, abrupt current transitions are observed at the backward and forward latch thresholds, with a low-current state maintained between the two transitions. Additional 1T-cSG characteristics, including device variation and endurance, are presented in Supplementary Note S2.6. It should be noted that without current compliance, the drain current continues to increase after latching because regenerative feedback is accompanied by enhanced channel conduction under the continuously increasing gate bias. Current compliance was therefore applied to prevent excessive current and stabilize the post-latch operating point. From the functional perspective of cSG, the compliance-limited current also provides a bounded output after threshold switching, suppressing the dependence of output amplitude on input magnitude. In a practical circuit, this current-limiting function can be readily implemented using a load or peripheral current-limiting circuitry. Furthermore, although the 1T-cSG is characterized in the current domain, its output can be converted into a voltage signal using a series resistor, enabling straightforward integration with voltage-mode circuitry (see Supplementary Note S2.7).

To verify that the bidirectional switching boundaries identified from the static characteristics can be selectively addressed during dynamic operation, pulse-based measurements were performed as shown in Fig. 2i. Starting from a standby V_{G} of 2.5 V, a positive-going V_{G} pulse to 1.5 V selectively induces the forward latch, whereas a negative-going pulse to 3.5 V induces the backward latch under $V_{\mathrm{D}} = 4$ V. The corresponding transient I_{D} responses exhibit distinct current pulses for the two latch events.

The measured responses confirm that the 1T-cSG can independently detect crossings of the upper and lower switching boundaries, effectively implementing a boxcar-filter function. During backpropagation, a surrogate-gradient signal of 1 is assigned when the 1T-cSG remains within the low-current region (i.e., the boxcar range), whereas a gradient of 0 is assigned outside the range. The bidirectional threshold response determines whether the synaptic input lies within or outside the surrogate-gradient window.

Co-integrated circuit operation on unit architecture

Having established the individual operation of the 1T-synapse, 1T-neuron, and 1T-cSG, we evaluate their integrated functionality within a unified hardware path interconnected through current-mirror interface circuits (see Methods, "Circuit implementations").

The 1T-synapse and 1T-neuron implement the forward path of the SNN system (Fig. 3a, left). The devices are interconnected through a breadboard-based current-mirror circuit composed of two commercial p-channel MOSFET (PMOS) transistors (See details in Supplementary Note S3.1). Presynaptic spikes applied to the 1T-synapse generate a synaptic current that is mirrored into the 1T-neuron and increases V_{D}, producing LIF neuronal dynamics (Fig. 3a, right). The incomplete reset of V_{mem} originates from the several-hundred-picofarad C_{par} of the breadboard setup. In a fully integrated implementation, C_{par} is expected to decrease below 100 fF, enabling V_{mem} to reset close to ground (See details in Supplementary Note S3.2).

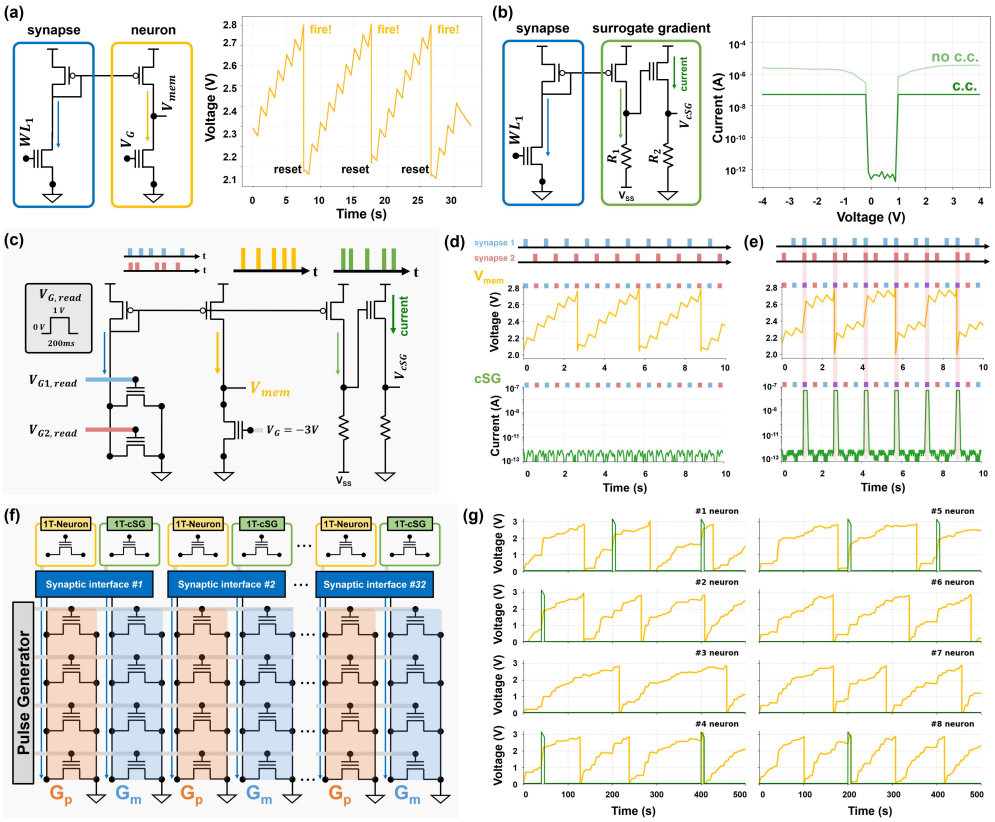


Fig. 3 Circuit-level evaluation of the co-integrated 1T architecture. (a) Circuit schematic of the co-integrated 1T-synapse and 1T-neuron connected through a current-mirror interface, and the measured LIF response. (b) Circuit schematic of the integrated 1T-synapse and a 1T-cSG, and the measured output current as a function of the word-line (WL_1) voltage. (c) Unit architecture comprising two 1T-synapses, one 1T-neuron, and one 1T-cSG. (d,e) Presynaptic raster plots, V_{mem}, and output current of the cSG for (d) non-overlapping and (e) overlapping input spikes. (f) Simulated 32×32 differential synaptic array with 32 1T-neurons and 32 1T-cSG modules. (g) Simulated membrane-potential and cSG responses of eight representative channels.

To extend the integrated hardware path from neuronal processing to state-access-free on-chip learning, the 1T-synapse is connected to the 1T-cSG through a current-mirror interface circuit (Fig. 3b, left). Fig. 3b, right, shows the measured cSG current as a function of the WL_1 voltage. At the standby WL_1 voltage of 0 V, the synaptic current biases the gate of the 1T-cSG within the low-current region. When the WL_1 voltage decreases below -0.2 V or increases above 1 V, the change in synaptic current drives V_{G} beyond the backward or forward latch threshold, respectively, producing a high cSG current. Accordingly, the 1T-cSG remains in the low-current state when the synaptic current lies within the boxcar range and switches to a high-current state outside the range.

The complete unit architecture integrates two 1T-synapses with a 1T-neuron and a 1T-cSG (Fig. 3c). When the two presynaptic spikes arrive at different synapses without temporal overlap, the 1T-cSG remains in the off state and the output current stays low, indicating that the input condition remains within the boxcar range (Fig. 3d). In contrast, temporal overlap between the two presynaptic spikes produces a high 1T-cSG output current, indicating that the input condition falls outside the boxcar range (Fig. 3e). The paired measurements demonstrate selective cSG responses to temporal coincidence while the neuron branch preserves LIF neuronal dynamics.

To evaluate system-level scalability, a single AIMC tile comprising a 32×32 AND-type differential 1T-synapse array, 32 1T-neurons, and 32 1T-cSG modules is evaluated using SPICE simulations based on a 180 nm process model (Fig. 3f). Each signed synaptic weight is encoded by a differential pair of devices with randomly programmed conductance states that generate diverse signed synaptic currents. Across all 32 channels, the neuron and cSG devices independently exhibit the intended LIF and cSG operations, supporting the scalability of the integrated block as a building unit for multi-tile neuromorphic systems (see Supplementary Note S3.3). The responses of eight representative channels (Fig. 3g) show neuronal membrane potentials evolving through integration of accumulated synaptic currents, while the corresponding 1T-cSG modules generate surrogate signals only when the synaptic currents remain within the prescribed operating window (see Supplementary Note S3.4). Consistent and independent responses across randomly programmed channels demonstrate the functional scalability of the integrated AIMC tile. SPICE simulations also validate circuit-level, spike-based temporal eligibility computation by the eLIFE module using 1T-neurons (see Supplementary Note S3.5).

Performance of training with state-access-free modules

Algorithm-level simulations evaluate cSG and eLIFE as the core state-access-free modules across networks and datasets of various scales (see Supplementary Note S4.1 for datasets, training settings, and network configurations). The performance of cSG is compared with that of conventional surrogate gradients (SG), while the effect of eLIFE is evaluated using otherwise matched configurations with and without eLIFE.

The small-scale evaluation considers two SNN architectures (Fig. 4a). Architecture 1 (Arch. 1) adopts a feedforward network topology, whereas Architecture 2 (Arch. 2) introduces feedback connections that provide an additional pathway for temporal information (see the ablation studies in Supplementary Note S4.3). Both architectures are trained using SLTT, which computes spatial gradients at every time step, or SOLO, which uses only the spatial gradient from the final time step.

The proposed state-access-free modules are first evaluated on small-scale networks (Fig. 4b). On the S-MNIST task, Arch. 1 reveals an algorithm-dependent effect of eLIFE. Adding eLIFE slightly lowers accuracy under SLTT but raises it under SOLO for both SG and cSG. On the SHD task, Arch. 1 shows a pronounced accuracy drop when cSG is used with SOLO. Incorporating eLIFE largely restores the lost accuracy, and eLIFE improves performance across all matched configurations on SHD. Across both temporal tasks, SOLO generally outperforms SLTT in Arch. 2, suggesting that the feedback architecture is well suited to settings that combine cSG with SOLO, in

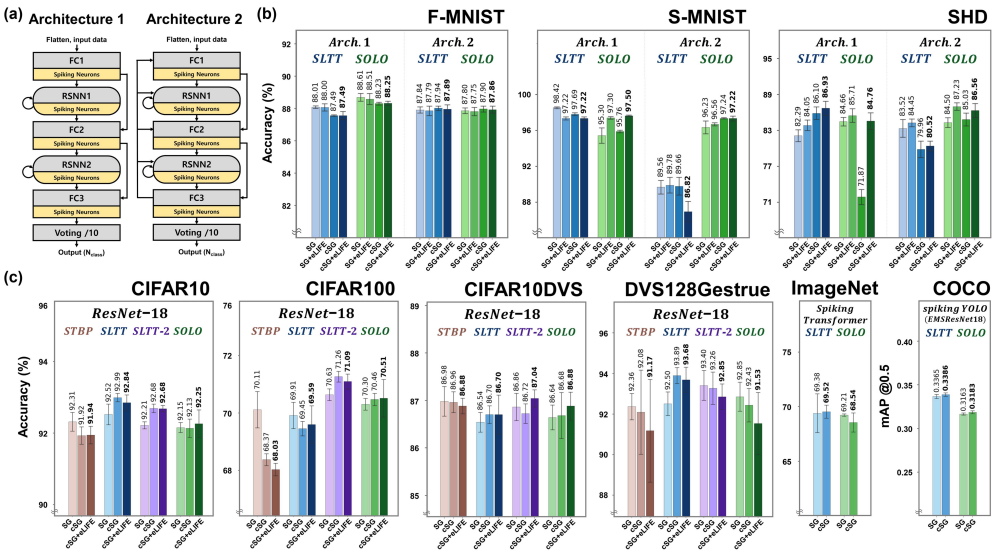


Fig. 4 Algorithm-level evaluation of state-access-free modules for hardware-friendly SNN training. (a) Network structures used for the small-scale evaluation. Arch. 1 uses feedforward inter-layer connections between recurrent spiking layers, whereas Arch. 2 adds feedback connections that provide an additional pathway for temporal information. (b) Classification accuracy on F-MNIST, S-MNIST, and SHD using SLTT or SOLO with SG or cSG and with eLIFE enabled or disabled. eLIFE produces algorithm-dependent changes on S-MNIST and restores accuracy after the pronounced cSG degradation under Arch. 1 with SOLO on SHD. (c) Large-scale evaluation using pre-activation ResNet-18 on CIFAR-10, CIFAR-100, CIFAR10DVS, and DVS128Gesture, Spikformer on ImageNet, and EMS-YOLO on COCO. Evaluated conditions include SG, cSG, and cSG with eLIFE, where applicable, under STBP, SLTT, SLTT-2, and SOLO. cSG and eLIFE generally preserve performance, particularly with spatial-gradient-based training.

which both the surrogate-gradient module and the training algorithm have limited access to temporal information.

Next, the large-scale evaluation encompasses static and event-based image classification and object detection using pre-activation ResNet-18, Spikformer, and EMS-YOLO (Fig. 4c). Pre-activation ResNet-18 is evaluated on the static CIFAR-10 and CIFAR-100 datasets and the event-based CIFAR10DVS and DVS128Gesture datasets. Spatio-temporal backpropagation (STBP) assigns credit across all time steps through both spatial and temporal gradients. Under STBP, replacing conventional SG with cSG, either alone or together with eLIFE, consistently lowers accuracy, with the largest reductions observed on CIFAR-100 and DVS128Gesture. Training with SLTT, SLTT-2, or SOLO relies exclusively on spatial gradients and often maintains or improves accuracy when cSG is used alone or together with eLIFE. Excluding STBP, cSG, either alone or together with eLIFE, does not degrade accuracy by more than 1.5% on any dataset evaluated with ResNet-18. The ResNet-18 results demonstrate that cSG and eLIFE can be incorporated into large-scale networks without substantial accuracy loss, particularly when training relies primarily on spatial-gradient computation.

Spikformer and EMS-YOLO are evaluated on ImageNet classification and COCO object detection, respectively[38,39]. On ImageNet, cSG does not degrade accuracy by more than 1% under either SLTT or SOLO relative to conventional SG. On COCO, cSG improves mAP by approximately 0.6% under both SLTT and SOLO. The ImageNet and COCO results support the scalability of cSG across large-scale classification and detection tasks.

Across the evaluated networks and datasets, cSG and eLIFE generally preserve performance while relying on hardware-accessible signals. Efficient training methods based on spatial gradients further reduce computational cost by avoiding temporal-gradient propagation, aligning state-access-free training with hardware-aware learning scenarios. The overall results support state-access-free training as a scalable alternative whose effectiveness depends on the network architecture and training algorithm.

Benchmark accuracy alone does not establish whether pretrained representations can be adapted under analog device constraints without degrading previously learned classes. Continual learning aims to acquire new information while retaining previously acquired knowledge[40].

Performance on analog on-chip continual learning

In real-world applications such as robotics, continual acquisition of new information is important as environments and operating conditions change over time[40]. The previous analog fine-tuning workflow[41] is extended to a simulated analog on-chip continual-learning framework. Measured characteristics of 1T-synapses, 1T-neurons, and 1T-cSG modules are incorporated into the network models to represent hardware nonidealities (See details in Supplementary Note S5.1). Additional spiking VGG9 experiments demonstrate that the framework remains effective at a larger scale under hardware nonidealities, achieving sufficient convergence within 10 epochs using the analog fine-tuning workflow (See details in Supplementary Note S5.2).

The simulated on-chip continual-learning framework is evaluated with sample-wise updates and one withheld target class per run (Fig. 5a). Selective weight updates and interleaved rehearsal, which revisits past training data to limit forgetting[42,43], are applied during adaptation. SOLO-pretrained Arch. 2 is adopted for online adaptation, with SLTT-pretrained initialization also evaluated (See details in Supplementary Note S5.3). Digital and analog synaptic implementations are compared to assess whether the energy-motivated transition to analog computation preserves algorithmic performance in online adaptation and retention under experimentally calibrated device characteristics. Within each implementation, cSG alone is compared with cSG combined with eLIFE to assess target-class learning and retention of previously seen classes.

The per-class accuracy-change maps show target-class learning with cSG and eLIFE even after incorporating analog characteristics (Fig. 5b,c). With cSG alone, substantial target-class gains are accompanied by pronounced losses in previously learned classes. The difference is particularly evident in the target-wise ten-class mean accuracy changes, which are positive for every target with eLIFE but strongly negative with cSG alone in both digital and analog implementations (Fig. 5d). Although cSG alone yields larger target-class gains (Fig. 5e), the accompanying old-class losses are much

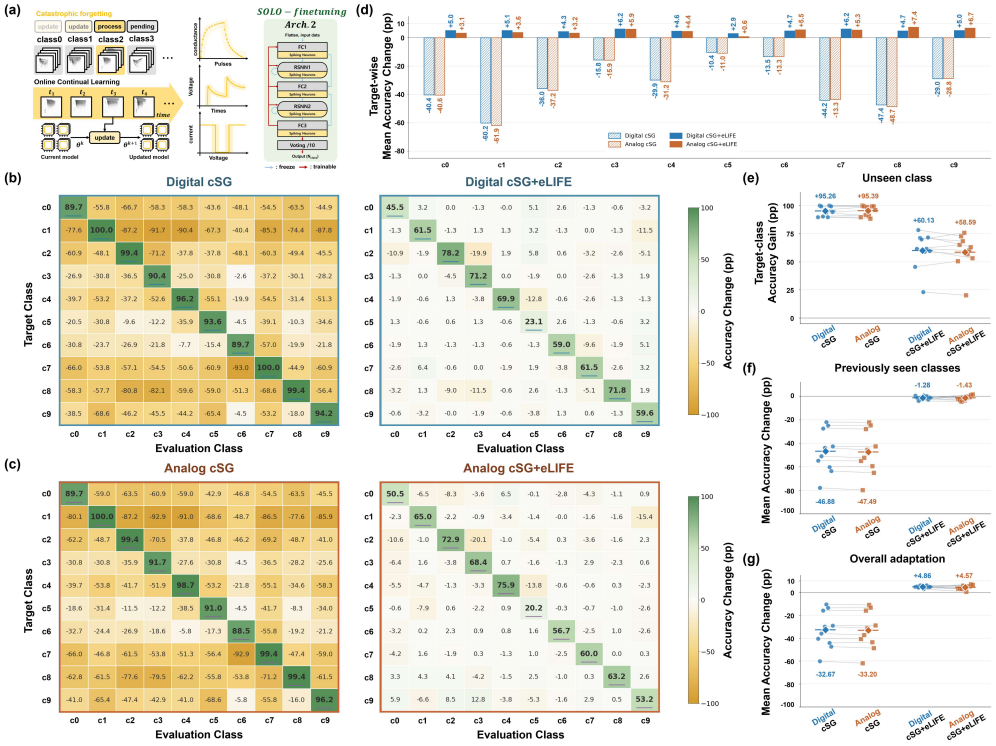


Fig. 5 Toward an analog on-chip continual-learning framework with state-access-free cSG and eLIFE modules. (a) Rehearsal-assisted online adaptation to one independently withheld class per run using SOLO-pretrained Arch. 2 weights, with one sample per weight update and selective weight training. (b,c) Per-class accuracy-change maps using cSG alone or with eLIFE for (b) digital and (c) analog synaptic implementations. Diagonal elements measure target-class accuracy gains and off-diagonal elements measure accuracy changes in previously seen classes. (d) Target-wise ten-class mean accuracy change, computed as the row-wise mean of each matrix in (b,c), showing the balance between target-class gains and old-class losses. (e–g) Digital–analog comparisons of (e) target-class accuracy gain, (f) mean accuracy change in previously seen classes, and (g) overall ten-class mean accuracy change. Connected points correspond to the same target class; diamonds and horizontal bars indicate means across targets.

greater than with eLIFE (Fig. 5f), resulting in negative rather than positive overall mean accuracy changes (Fig. 5g). Under the tested adaptation conditions, combining cSG with eLIFE improves the balance between unseen-class learning and retention of previously seen classes. Reduced old-class losses are also observed when eLIFE is added to conventional digital SG, indicating that the retention benefit is not limited to the current-based surrogate (See details in Supplementary Note S5.4). Exploratory t-SNE and accumulate-neuron analyses further examine feature representations and output-state dynamics before and after adaptation (See details in Supplementary Note S5.5).

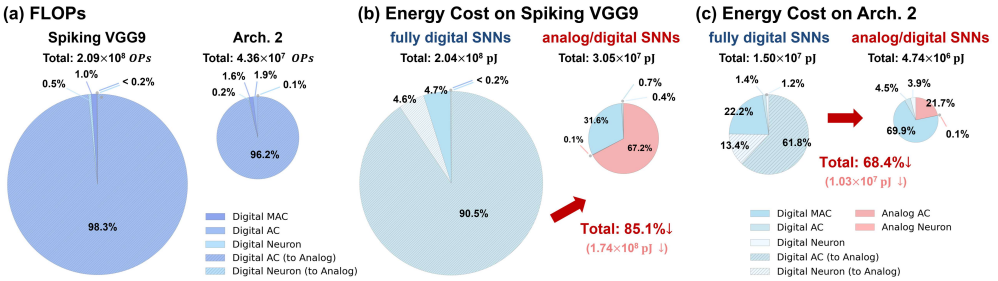


Fig. 6 Model-based inference-energy estimates for analog SNNs. (a) FLOPs evaluation for Spiking VGG-9 and Arch. 2 using digital-equivalent operation counts and mapping of synaptic operations to analog accumulation (AC). (b) Estimated inference energy for Spiking VGG-9 under the adopted model, decreasing by 85.05% relative to the fully digital baseline, from 2.04×10^8 to 3.05×10^7 pJ, a modeled saving of 1.735×10^8 pJ. (c) Estimated inference energy for Arch. 2, decreasing by 68.4% under the same model, from 1.50×10^7 to 4.74×10^6 pJ, a modeled saving of 1.026×10^7 pJ.

Discussion

We present a hardware–algorithm co-designed neuromorphic architecture for state-access-free learning and inference. The cSG and eLIFE modules avoid external access to internal membrane potentials during gradient computation. Fabricated devices and the measured 1T-based unit architecture establish the physical feasibility of synaptic storage, neuronal dynamics, and cSG generation. SPICE simulations support eLIFE circuit operation, while network simulations incorporating measured device characteristics demonstrate a hardware-compatible analog fine-tuning workflow with competitive algorithmic performance.

The potential inference-energy benefit of the analog implementation is assessed for spiking VGG-9 and Arch. 2 using digital-equivalent FLOPs and an energy model that includes basal neuronal updates and event-driven synaptic and neuronal activity (Fig. 6a). Mapping a substantial fraction of the high-density network operations to analog accumulation replaces energy-intensive digital cycles. Under the adopted model and energy parameters, the analog-dominant implementations reduce total inference energy by 85.05% for spiking VGG-9 and 68.4% for Arch. 2 relative to fully digital baselines, corresponding to savings of 1.735×10^8 and 1.026×10^7 pJ per inference, respectively (Fig. 6b,c). The estimated reductions arise mainly from the small modeled contribution of analog neuronal dynamics and the lower modeled synaptic-energy cost of analog accumulation (AC).

Training energy requires further investigation. At the device level, high gate voltages trigger Fowler-Nordheim (FN) tunneling while the source and drain terminals remain grounded, suppressing channel-current flow during programming. In a practical system, however, high-voltage operation requires charge-pump circuits that can consume several milliwatts (mW), underscoring the need for chip-level optimization of training energy in a fully integrated implementation[44]. Selective fine-tuning of the feedback and output weights may reduce costly programming operations relative to training from scratch, but programming counts have not been measured. The energy, area, and latency of a complete integrated learning system have not been evaluated.

Homogeneous co-integration of 1T-neurons, 1T-synapses, and 1T-cSGs validates the common device platform and the physical feasibility of the proposed functions. The current-mirror interfaces used in the unit demonstration were implemented with external components; monolithic integration of equivalent peripheral circuits with 1T-devices and control logic remains a prospective route toward system scalability. Compared with circuit-based spiking neurons, the proposed 1T-neuron provides a compact footprint as small as $6F^2$, supporting the high-density integration required for scalable neuromorphic systems (see details in Supplementary Table S5). Supplementary Table S6 compares the 1T-neuron with representative device-based neuron implementations.

Taken together, these results support formulating learning rules around hardware-accessible signals. The observed architecture- and training-rule-dependent outcomes delimit the tested scope of this approach. Further comparisons with hardware-friendly local learning rules[36] and shared neuromorphic benchmarks[45] would help identify the operating regimes in which such co-design is most beneficial.

Methods

Preliminaries

Leaky integrate-and-fire neuron model

The LIF neuron model is adopted because its simple dynamics are well suited to spiking-network simulation and analog implementation[46]. The subthreshold membrane-potential dynamics are defined as

$$\tau_{\mathrm{mem}} \frac{\partial V_{\mathrm{mem}}(t)}{\partial t} = -(V_{\mathrm{mem}}(t) - V_{\mathrm{reset}}) + I(t), \tag{1}$$

where $V_{\mathrm{mem}}(t)$ is the subthreshold membrane potential, V_{reset} is the reset potential after spike generation, τ_{mem} is the membrane time constant, and $I(t)$ is the normalized synaptic input.

Discrete leaky integrate-and-fire neuron model

The discrete computational form of the LIF neuron model is defined as

$$I_i^l[t] = \sum_j W_{ij}^l S_j^{l-1}[t] \tag{2}$$

$$V_i^l[t] = \beta V_i^l[t-1](1 - S_i^l[t-1]) + I_i^l[t] \tag{3}$$

$$S_i^l[t] = \Theta(V_i^l[t] - \vartheta_{\mathrm{th}}), \tag{4}$$

where $I_i^l[t]$ is the synaptic input to neuron i in layer l, W_{ij}^l is the weight from presynaptic neuron j to postsynaptic neuron i, and $S_j^{l-1}[t]$ is the binary presynaptic spike. The variables $V_i^l[t]$ and $S_i^l[t]$ denote the membrane potential and output spike of neuron i, respectively. The membrane-decay factor is $\beta = \exp(-\Delta t/\tau_{\mathrm{mem}})$ for simulation

time step Δt, $\Theta(\cdot)$ is the Heaviside function, and ϑ_{th} is the firing threshold. Equation 4 assumes $V_{\text{reset}} = 0$, no bias current, and a multiplicative hard reset.

Surrogate gradient

Owing to the non-differentiability of the Heaviside function, the surrogate gradient method is employed[27]. In particular, the boxcar function is adopted because of its suitability for efficient on-chip implementation.

$$\frac{\partial S_i^l[t]}{\partial V_i^l[t]} \approx \mathbb{1}_{V,i}^l[t] = \Theta\big(p_V - \big|V_i^l[t] - \vartheta_{\text{th}}\big|\big)\,, \tag{5}$$

where $\mathbb{1}_{V,i}^l[t]$ is the binary boxcar output and p_V is the half-width of the voltage-domain surrogate-gradient window.

State-access-free modules

Current-based surrogate gradient

The current-based surrogate gradient addresses the inaccessibility of membrane potentials by deriving the binary surrogate signal from the synaptic current. The current-domain boxcar is defined as

$$\mathbb{1}_{I,i}^l[t] = \Theta\big(p_I - \big|I_i^l[t] - I_{\text{th}}\big|\big)\,, \tag{6}$$

where I_{th} and p_I are the center and half-width of the current-domain cSG window, respectively. During gradient computation, $\mathbb{1}_{I,i}^l[t]$ replaces the voltage-dependent term $\mathbb{1}_{V,i}^l[t]$. The low-current state of the 1T-cSG corresponds to $\mathbb{1}_{I,i}^l[t] = 1$, whereas a latch event outside the prescribed window corresponds to $\mathbb{1}_{I,i}^l[t] = 0$.

Eligible leaky integrate-and-fire encoder

To implement eligibility traces on neuromorphic chips without exposing continuous states, an eLIFE module is associated with each encoded spike stream. The module uses two interconnected clone neurons with LIF dynamics. Let $S_{m,i}^l[t]$ denote the main-neuron spike supplied to the eLIFE module associated with neuron i in layer l. The clone-neuron dynamics are

$$V_{c1,i}^l[t] = \beta_{c1} V_{c1,i}^l[t-1](1 - S_{c1,i}^l[t-1]) + S_{c2,i}^l[t-1] + S_{m,i}^l[t] \tag{7}$$

$$S_{c1,i}^l[t] = \Theta(V_{c1,i}^l[t] - \vartheta_{\text{th}}) \tag{8}$$

$$V_{c2,i}^l[t] = \beta_{c2} V_{c2,i}^l[t-1](1 - S_{c2,i}^l[t-1]) + S_{c1,i}^l[t-1] \tag{9}$$

$$S_{c2,i}^l[t] = \Theta(V_{c2,i}^l[t] - \vartheta_{\text{th}}), \tag{10}$$

where $V_{c1,i}^l[t]$ and $V_{c2,i}^l[t]$ are the membrane potentials of clone neurons $c1$ and $c2$, and $S_{c1,i}^l[t]$ and $S_{c2,i}^l[t]$ are the corresponding binary spikes. The deployed clone-neuron decay factors are $\beta_{c1} = 0.5$ for the clone receiving the main-neuron spike and

$\beta_{c2} = 1$ for the recurrently driven clone. The input and cross-coupling weights in Eq. 10 are fixed at unity, and no sigmoid is applied to the fixed coefficients during network execution. The Heaviside function and firing threshold follow the definitions used for the main LIF neuron (See details in Supplementary Note S4.2).

The eligible spike is the logical OR of the main- and clone-neuron spikes:

$$\hat{S}_i^l[t] = S_{m,i}^l[t] \vee S_{c1,i}^l[t] \vee S_{c2,i}^l[t], \tag{11}$$

where $\hat{S}_i^l[t]$ is the final eligible spike associated with neuron i at layer l.

Device fabrication

For hardware implementation of the analog SNN system, homogeneous 1T-based neurons, synapses, and cSG modules were fabricated on a single 8-inch p-type (100) silicon-on-insulator (SOI) wafer with a 145 nm buried-oxide (BOX) layer and a 55 nm top-Si layer. The active regions were patterned first, followed by deposition of an ONO stack composed of 3 nm SiO_2, 6 nm Si_3N_4, and 8 nm SiO_2. A 100 nm heavily doped n^+ polycrystalline-Si layer was then deposited and patterned to form the gates. Finally, self-aligned phosphorus implantation and rapid thermal annealing were performed to form the source and drain regions.

Circuit implementations

LIF 1T-neuron and 1T-synapse

The circuit was operated at $V_{\mathrm{DD}} = 3.3$ V. Presynaptic pulses with an amplitude of 1 V and a duration of 200 ms were applied to the gate of the 1T-synapse, pulling down V_{G} of the current-injecting PMOS transistor connected to the drain of the 1T-neuron. The resulting 200 ms current injection raised V_{mem} during integration, followed by a 300 ms leakage interval after the presynaptic pulse. When V_{mem} reached V_{firing}, the 1T-neuron fired and reset V_{mem}.

1T-cSG and 1T-synapse

V_{DD} and V_{SS} were set to 3.3 V and -4 V, respectively, to establish the operating range of the 1T-cSG. At sufficiently low WL_1 voltage, the synapses provide insufficient current to raise the 1T-cSG gate voltage, and BTBT maintains the device in the backward-latched state. Increasing WL_1 raises the synaptic current mirrored to the 1T-cSG gate. The backward latch is released when V_{G} crosses the lower threshold, placing the device in the high-resistance state (HRS). A further increase in WL_1 triggers the I.I.-induced forward latch.

Unit architecture

In the unit architecture, a resistive divider between V_{DD} and V_{SS} with a 1:4 ratio biased the 1T-cSG gate near -2.5 V. The device remained in the HRS in the absence of presynaptic spikes. When two presynaptic pulses with an amplitude of 1 V and a duration of 200 ms overlapped, the two 1T-synapses supplied sufficient current to raise V_{G}, triggering the forward-latched state and producing a high output current.

Algorithm-level evaluation configuration

Algorithm-level evaluation compares SG and cSG, with eLIFE enabled or disabled as an ablation variable, under matched network, optimizer, and training-rule settings.

Continual-learning performance analysis

For each independently withheld SHD class t, accuracy changes are measured from the condition-specific pre-adaptation baseline to the end of all 20 target-learning/rehearsal rounds. The class-wise accuracy change is

$$\Delta A_{t,j} = A_{t,j}^{\text{post}} - A_{t,j}^{\text{base}}, \tag{12}$$

where j indexes Classes 0–9. Diagonal entries measure target-class gain, while off-diagonal entries measure changes in previously seen classes. The overall ten-class accuracy change is the row-wise mean,

$$\Delta\text{ACC}_t = \frac{1}{10}\sum_{j=0}^{9} \Delta A_{t,j}. \tag{13}$$

Calculation of energy cost

The total energy consumption of the proposed SNN architecture, E_{total}, is modeled as the sum of the energy dissipated in digital layers, $\mathbb{L}_d$, and analog layers, $\mathbb{L}_a$. For each layer, the energy is decomposed into synaptic and neuronal components and accumulated over T simulation time steps (See details in Supplementary Note S6.2).

$$\begin{aligned} E_{\text{total}} = &\sum_{l\in\mathbb{L}_d} T\cdot\Big[\underbrace{\left(OP_{MAC}^{(l)}E_{MAC}^{d} + f_{r,in}^{(l)}OP_{AC}^{(l)}E_{AC}^{d}\right)}_{\text{digital synaptic energy}} + \underbrace{N_{\text{neu}}^{(l)}\left(E_{up}^{d} + f_{r,out}^{(l)}E_{fire}^{d}\right)}_{\text{digital neuronal energy}}\Big] \\ &+ \sum_{l\in\mathbb{L}_a} T\cdot\Big[\underbrace{\left(2f_{r,in}^{(l)}OP_{AC}^{(l)}E_{AC}^{a}\right)}_{\text{analog synaptic energy}} + \underbrace{N_{\text{neu}}^{(l)}\left(E_{up}^{a} + f_{r,out}^{(l)}E_{fire}^{a}\right)}_{\text{analog neuronal energy}}\Big], \end{aligned} \tag{14}$$

Here, $OP_{MAC}^{(l)}$ and $OP_{AC}^{(l)}$ are the numbers of multiply-accumulate (MAC) and accumulate (AC) operations mapped to layer l per time step, and $N_{\text{neu}}^{(l)}$ is the number of neurons in the layer. The dimensionless quantities $f_{r,in}^{(l)}$ and $f_{r,out}^{(l)}$ are the average fractions of input and output neurons that spike in one time step. The parameters E_{MAC}^{d} and E_{AC}^{d} are the energy costs per digital MAC and digital AC, and E_{AC}^{a} is the energy cost per individual 1T synapse. The factor of 2 in the analog synaptic term accounts for the two physical synapses in the differential $G_{\text{P}}/G_{\text{M}}$ pair used to represent each signed weight. The basal update energies per neuron and time step are E_{up}^{d} and E_{up}^{a}, while E_{fire}^{d} and E_{fire}^{a} are the dynamic energies dissipated per spike.

For the digital synaptic energy parameters, E^d_{MAC} is 4.6 pJ and E^d_{AC} is 0.9 pJ[47]. The analog value E^a_{AC} is estimated as 0.05 pJ per individual 1T synapse from the interface-inclusive read power and pulse duration. The explicit factor of two in Eq. 14 therefore yields 0.10 pJ per logical signed operation represented by a differential $G_{\mathrm{P}}/G_{\mathrm{M}}$ pair. The adopted digital neuronal energy parameters are $E^d_{up} = 9.736$ pJ and $E^d_{fire} = 1.0$ pJ[48], whereas the analog neuronal values $E^a_{up} = 0.0072$ pJ and $E^a_{fire} = 0.3136$ pJ are estimated from the capacitance-based model. The high-level operation counts $OP^{(l)}_{MAC}$ and $OP^{(l)}_{AC}$ incorporate hardware-specific mapping factors, including the number of processing-element cycles, $C^{(l)}_{pe}$, and the 32×32 crossbar dimension, N_{xbar}.

Data availability

The public datasets used in this study are available from their respective sources, including MNIST, Fashion-MNIST, CIFAR-10, CIFAR-100, CIFAR10-DVS, DVS128Gesture, SHD, ImageNet-1K, and COCO 2017. The generated data supporting the device-characterization measurements, network simulations, continual-learning analyses, and energy estimates can be requested from the corresponding authors. The limitations of the retained aggregate continual-learning records, including the inability to reconstruct all paired seed-wise variances, are described in the Supplementary Information.

Code availability

The custom code used for SNN training, implementation of the cSG and eLIFE modules, analog-aware post-training, tile-level simulation analysis, and energy estimation will be made available to editors and reviewers upon reasonable request during peer review. The code will be publicly released upon acceptance.

References

1. Kudithipudi, D. *et al.* Neuromorphic computing at scale. *Nature* **637**, 801–812 (2025).

2. Maass, W. Networks of spiking neurons: The third generation of neural network models. *Neural Netw.* **10**, 1659–1671 (1997).

3. Merolla, P. A. *et al.* A million spiking-neuron integrated circuit with a scalable communication network and interface. *Science* **345**, 668–673 (2014).

4. Pei, J. *et al.* Towards artificial general intelligence with hybrid Tianjic chip architecture. *Nature* **572**, 106–111 (2019).

5. Davies, M. *et al.* Loihi: A neuromorphic manycore processor with on-chip learning. *IEEE Micro* **38**, 82–99 (2018).

6. Orchard, G. *et al.* Efficient neuromorphic signal processing with Loihi 2. In *2021 IEEE Workshop on Signal Processing Systems (SiPS)*, 254–259 (IEEE, Coimbra, Portugal, 2021).

7. Furber, S. B., Galluppi, F., Temple, S. & Plana, L. A. The SpiNNaker project. *Proc. IEEE* **102**, 652–665 (2014).

8. Gonzalez, H. A. *et al.* SpiNNaker2: A large-scale neuromorphic system for event-based and asynchronous machine learning. Preprint at https://arxiv.org/abs/2401.04491 (2024).

9. Schemmel, J. *et al.* A wafer-scale neuromorphic hardware system for large-scale neural modeling. In *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*, 1947–1950 (IEEE, Paris, France, 2010).

10. Pehle, C. *et al.* The BrainScaleS-2 accelerated neuromorphic system with hybrid plasticity. *Front. Neurosci.* **16**, 795876 (2022).

11. Kumar, S., Wang, X., Strachan, J. P., Yang, Y. & Lu, W. D. Dynamical memristors for higher-complexity neuromorphic computing. *Nat. Rev. Mater.* **7**, 575–591 (2022).

12. Aguirre, F. *et al.* Hardware implementation of memristor-based artificial neural networks. *Nat. Commun.* **15**, 1974 (2024).

13. Roy, K., Jaiswal, A. & Panda, P. Towards spike-based machine intelligence with neuromorphic computing. *Nature* **575**, 607–617 (2019).

14. Wang, J., Shi, Y., Chen, Z. & Wen, M. SpikeFlow: A hardware–software co-designed systolic array for spiking neural networks. *J. Syst. Archit.* **169**, 103604 (2025).

15. Zhang, J. *et al.* ANP-I: A 28-nm 1.5-pJ/SOP asynchronous spiking neural network processor enabling sub-0.1-μJ/sample on-chip learning for edge-AI applications. *IEEE J. Solid-State Circuits* **59**, 2717–2729 (2024).

16. Xia, Q. & Yang, J. J. Memristive crossbar arrays for brain-inspired computing. *Nat. Mater.* **18**, 309–323 (2019).

17. Joubert, A., Belhadj, B., Temam, O. & Heliot, R. Hardware spiking neurons design: Analog or digital? In *The 2012 International Joint Conference on Neural Networks (IJCNN)*, 1–5 (IEEE, Brisbane, Australia, 2012).

18. Indiveri, G., Chicca, E. & Douglas, R. A VLSI array of low-power spiking neurons and bistable synapses with spike-timing dependent plasticity. *IEEE Trans. Neural Netw.* **17**, 211–221 (2006).

19. Aamir, S. A. *et al.* A mixed-signal structured AdEx neuron for accelerated neuromorphic cores. *IEEE Trans. Biomed. Circuits Syst.* **12**, 1027–1037 (2018).

20. Han, J.-K. *et al.* 3D stackable broadband photoresponsive InGaAs biristor neuron for a neuromorphic visual system with near 1 V operation. In *2021 IEEE International Electron Devices Meeting (IEDM)*, 1–4 (IEEE, San Francisco, CA, USA, 2021).

21. Han, J.-K. *et al.* Mimicry of excitatory and inhibitory artificial neuron with leaky integrate-and-fire function by a single MOSFET. *IEEE Electron Device Lett.* **41**, 208–211 (2020).

22. Han, C. Y. *et al.* Configurable NbO_x memristors as artificial synapses or neurons achieved by regulating the forming compliance current for the spiking neural network. *Adv. Electron. Mater.* **9**, 2300018 (2023).

23. Yuan, R. *et al.* A calibratable sensory neuron based on epitaxial VO2 for spike-based neuromorphic multisensory system. *Nat. Commun.* **13**, 3973 (2022).

24. Lee, J.-H., Yun, S.-Y., Han, J.-K. & Choi, Y.-K. A review of single transistor-based neuron for highly scalable neuromorphic hardware. *IEEE Electron Devices Rev.* **2**, 188–200 (2025).

25. Han, J.-K. *et al.* Cointegration of single-transistor neurons and synapses by nanoscale CMOS fabrication for highly scalable neuromorphic hardware. *Sci. Adv.* **7**, eabg8836 (2021).

26. Pazos, S. *et al.* Synaptic and neural behaviours in a standard silicon transistor. *Nature* **640**, 69–76 (2025).

27. Neftci, E. O., Mostafa, H. & Zenke, F. Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking

neural networks. *IEEE Signal Process. Mag.* **36**, 51–63 (2019).

28. Wu, Y., Deng, L., Li, G., Zhu, J. & Shi, L. Spatio-temporal backpropagation for training high-performance spiking neural networks. *Front. Neurosci.* **12**, 331 (2018).

29. Bellec, G. *et al.* A solution to the learning dilemma for recurrent networks of spiking neurons. *Nat. Commun.* **11**, 3625 (2020).

30. Xiao, M., Meng, Q., Zhang, Z., He, D. & Lin, Z. Online training through time for spiking neural networks. In *Advances in Neural Information Processing Systems*, Vol. 35, 20717–20730 (Curran Associates, Inc., 2022). URL https://proceedings.neurips.cc/paper_files/paper/2022/file/82846e19e6d42ebfd4ace4361def29ae-Paper-Conference.pdf.

31. Meng, Q. *et al.* Towards memory- and time-efficient backpropagation for training spiking neural networks. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 6143–6153 (IEEE, Paris, France, 2023).

32. Lee, S., Moon, H., Yoon, M., Heo, W. & Kim, S. SOLO: Spatial online learning at once for spiking neural networks. *Neurocomputing* **663**, 131940 (2026).

33. Zenke, F. & Neftci, E. O. Brain-inspired learning on neuromorphic substrates. *Proc. IEEE* **109**, 935–950 (2021).

34. Wan, W. *et al.* A compute-in-memory chip based on resistive random-access memory. *Nature* **608**, 504–512 (2022).

35. Fu, H. *et al.* SpikeRAM: A 48.1pW/synapse/bit event-driven spiking compute-near/in-memory processor with neuromorphic sensor enabling life-long on-chip learning. In *2026 IEEE International Solid-State Circuits Conference (ISSCC)*, 314–316 (IEEE, San Francisco, CA, USA, 2026).

36. Ororbia, A. G. Brain-inspired machine intelligence: A survey of neurobiologically-plausible credit assignment. Preprint at https://arxiv.org/abs/2312.09257 (2023).

37. Chen, C.-D. *et al.* Single-transistor latch in SOI MOSFETs. *IEEE Electron Device Lett.* **9**, 636–638 (1988).

38. Zhou, Z. *et al.* Spikformer: When spiking neural network meets transformer. Preprint at https://arxiv.org/abs/2209.15425 (2022).

39. Su, Q. *et al.* Deep directly-trained spiking neural networks for object detection. Preprint at https://arxiv.org/abs/2307.11411 (2023).

40. Parisi, G. I., Kemker, R., Part, J. L., Kanan, C. & Wermter, S. Continual lifelong learning with neural networks: A review. *Neural Netw.* **113**, 54–71 (2019).

41. Kang, J. *et al.* Achieving low-power analog resistive switching in filamentary memristive devices for energy-efficient analog in-memory computing. *Small* **21**, e05708 (2025).

42. Rebuffi, S.-A., Kolesnikov, A., Sperl, G. & Lampert, C. H. iCaRL: Incremental classifier and representation learning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5533–5542 (IEEE, Honolulu, HI, 2017).

43. Rolnick, D., Ahuja, A., Schwarz, J., Lillicrap, T. & Wayne, G. Experience replay for continual learning. In *Advances in Neural Information Processing Systems*, Vol. 32 (Curran Associates, Inc., 2019). URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/fa7cdfad1a5aaf8370ebeda47a1ff1c3-Abstract.html.

44. Xu, T. *et al.* A programmable high efficiency charge pump system for embedded flash memory with improved current driving capability. *Microelectron. J.* **159**, 106626 (2025).

45. Yik, J. *et al.* The neurobench framework for benchmarking neuromorphic computing algorithms and systems. *Nat. Commun.* **16**, 1545 (2025).

46. Gerstner, W. & Kistler, W. M. *Spiking Neuron Models: Single Neurons, Populations, Plasticity* (Cambridge University Press, Cambridge, 2002).

47. Horowitz, M. 1.1 Computing's energy problem (and what we can do about it). In *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, 10–14 (IEEE, San Francisco, CA, USA, 2014).

48. Moitra, A. *et al.* SpikeSim: An end-to-end compute-in-memory hardware evaluation tool for benchmarking spiking neural networks. *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.* **42**, 3815–3828 (2023).

Acknowledgements

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2024-00466640, "Development and Verification of an Ultra-Low Power PIM Chip Based on a Novel IGZO TFT Synaptic Device Co-designed and Integrated with a Commercial Foundry Si CMOS Platform"), the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2025-02263706, "Development of an Analog-Digital Mixed Ultra-Low-Power Neuromorphic Edge SoC"), and K-CHIPS (Korea Collaborative & High-tech Initiative for Prospective Semiconductor Research) (RS-2025-02305568, 25064-15TC) funded by the Ministry of Trade, Industry & Energy (MOTIE, Korea).

Author contributions

S.L. and J.-H.L. designed the study concept. S.L., W.H., and T.H. built the algorithmic simulation framework and conducted the model training. S.L. developed the analog on-chip fine-tuning workflow. S.L. and J.-H.L. performed the energy estimation. J.-K.H. fabricated the device. J.-H.L. and S.C.B. performed the electrical measurements. J.-Y.P., M.-H.J., and S.P. analyzed the measured data. S.L., J.-H.L., S.K., and J.-K.H. wrote the manuscript. S.K. and J.-K.H. supervised the project, secured funding, and provided overall direction. All authors discussed the results and commented on the manuscript at all stages.

Competing interests

The authors declare no competing interests.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation.pdf