Electronics, Communications, and Computing Summit
Vol. 2, No. 2, Apr - Jun 2024, pp. 51-57
DOI: https://doi.org/10.17051/ECC/02.02.07

Design and Evaluation of Neuromorphic Hardware Architectures for Low-Power Edge AI Applications

Zhiyi Chen[1], Q. Hugha[2]

[1]The Third Affiliated Hospital of Guangzhou Medical University, China, Email: winchen@vip.126.com
[2]Robotics and Automation Laboratory Universidad Privada Boliviana Cochabamba, Bolivia.

Article Info

Article history:

Received : 17.04.2024
Revised : 19.05.2024
Accepted : 21.06.2024

Keywords:

Neuromorphic computing,
Spiking Neural Networks (SNN),
Edge AI,
Low-power hardware,
FPGA,
Memristor,
STDP,
Event-driven processing

ABSTRACT

The paper focuses on designing and testing neuromorphic systems in hardware, and in particular those designed to perform edge AI systems with low power and real-time performance in mind. The main aim is to design Event-driven computing systems that emulate neural processes of nature to obtain energy-efficient inference at the edge. A CMOS-based and memristor-based paradigm of spiking neural network (SNN) accelerators are discussed. Leaky integrate-and-fire (LIF) neurons and leaky integrate-and-fire (LIF) neurons as well as synaptic integration instances are mapped to FPGA platforms as modular RTL implementations ready to be exploited and benchmarked in this FPGA-based prototyping workflow. Edge-relevant tasks optimal to the proposed neuromorphic cores include handwritten digit classification and dynamic vision-based gesture recognition and voice command detection. Power-performance comparing with the traditional multiply-accumulate (MAC) based AI accelerators is discussed. There are up to 70 percent dynamic power reduction and a 3x factor improvement in energy-per-inference has been observed, reflecting the architectural compatibility of SNNs to constrained edge environments. More so, this paper reviews trade-offs of on-chip learning flexibility, inversion latency, and hardware extensiveness. A deployment model at system level is proposed to provide an example of integration in the real world into edge AI stacks with emphasis on the modularity of interaction between neuromorphic processing components and embedded components. The results confirm that neuromorphic architectures bring very strong benefits to edge applications and especially in latency-, energy-, and area-sensitive applications. The scale of such a system, as well as the related design considerations, is an additional topic of interest raised in the work as it provides insight into future directions of ASIC design and adaption to hybrid edge-AI pipelines.

1. INTRODUCTION

The high rate of edge computing devices coming into various fields like smart wearables, autonomous sensors and IoT systems has increased the necessity of low latency, energy efficient, artificial intelligence (AI) on the edge. Traditional AI accelerators that are often digital von Neumann systems are highly dependent on dense multiply-accumulate (MAC) functions that are power-hungry and not performant to tighter energy and low-latency demands at the edge. They are memory bottlenecks too because computing and storing have been separated and this increases the amount of energy to move the data and restricts real time responsiveness. Neuromorphic computing is an alternative inspired by biology, whereby computation is event-based whereby each computation tightly interacts with memory, emulating the structure and dynamics of the human brain. The fundamental computational model used in neuromorphic systems is Spiking Neural Networks (SNNs), and they compute information using sparse, spike-based signaling, which allow truly ultra-low power utilize, and they are inherently asynchronous. Existing successful prototyping of hardware (like the Loihi (Intel, 2019) or TrueNorth (Merolla et al., 2014) systems) have proven attractive energy characteristics, yet are too advanced (or domain-specific) to scale to real-world edge AI applications (Davies et al., 2018). Nevertheless, existing neuromorphic hardware is not modular, not reconfigurable and has not been benchmarked under realistic edge workloads. In addition, architectural complexity, power efficiency, and inference accuracy trade-offs in constrained settings are under-studied.

Contributions that fill these gaps are characterizing and evaluating 2 RTL-level implementations of

neuromorphic hardware, memristor- and CMOS-based accelerators of SNNs focused on edge AI applications, designed in this paper. Our evaluation on the use of FPGA prototyping and workload simulation on power, latency and scalability leads to the provision of a practical roadmap on the future ASIC integration into edge AI applications. The described neuromorphic hardware is especially well-suited to real-time edge AI applications like wearable health devices, autonomous environmental monitoring, localable speech or gesture recognition systems, and so on, where low power, on-chip learning, and very-low latency lifetime requirements are paramount.

2. RELATED WORK

A number of neuromorphic hardware systems have more recently been developed to simulate brain-like computation using spiking neural networks (SNNs) and event-driven systems. Markedly, Intel Loihi (Davies et al., 2018) introduces programmable SNN brains on a chip with on-chip learning, providing energy-efficient real-time inference to execute a robotic and cognitive processing. A massively parallel neurosynaptic chip with more than one million neurons and 256 million synapses was created by IBM TrueNorth (Merolla et al., 2014) that is extremely energy-efficient at pattern recognition. Equally, BrainScaleS (Schemmel et al., 2020) applies the analog VLSI circuit to the SNN emulation at high-speed rates and hybrid learning-experiments, especially in neuroscience studies. These platforms are mostly architected in centralized or lab scale settings though they have significant architectural innovations and do not scale well to the edge setting. They can be code-dependent, chip-sized, non-industry-standard low-power interface and customization options. Moreover, they are complex and therefore hard to integrate into resource-constrained systems at the edges especially when real-time guarantees, power-restricted budgets, task-specific reconfigurabilities are needed. Furthermore, there are a few reports of comparative architectural assessment in realistic edge workloads. There is little systematic comparison of power/latency/inference accuracy of SNN designs over microcontroller-like or FPGA-level hardware. The absence of common metrics and publicly available architectural foundations in neuromorphic computing decreases the feasibility of using it in the development of edge AI. In this paper, we overcome these shortcomings through the design, benchmarking of two RTL based neuromorphic accelerator, study trade offs in energy, area, and performance at edge-relevant scenario, and propose a path towards a scalable design roadmap of low power neuromorphic integration in future.

3. Neuromorphic Architecture Design

Neuromorphic systems apply the structural and functional dynamics of biological neural systems, to achieve sparse, asynchronous, energy-efficient computation. The architecture to be proposed rests on the achievement of integrating: Leaky Integrate-and-Fire (LIF) neurons and Spike-Timing Dependent Plasticity (STDP - based) synaptic learning that are modeled at the Register Transfer Level (RTL) in FPGA implementation as well as ASIC aimed.

3.1 LIF Neuron Model

LIF neurons imitate the temporal integration process of multiple input spikes and it is characteristic of LIF to behave dynamically on thresholding levels. At some point when the summed up membrane potential reaches a threshold, the neuron fires and is reset just like the biological firing behavior. Refractory periods and configurable decay constants are supported with the model suitable to be used in low-complexity hardware designs. FPGA update logic is fixed-point arithmetic with clock-synched updates with each neuron.

3.2 STDP Synapse Design

STDP Synapses Bio-realistic synaptic learning Hebbian theory of learning is a process of changing the strength of connections (synaptic weights) between neurons as a result of the time relationship between a pre-synaptic action potential and a post-synaptic one. In hardware, as implemented by Hess, STDP is implemented with a looked up table and event-logging registers recording spike timestamps and adjusting the weight accordingly. This employs local learning without centralised oversight and is a primitve technology for localised on-chip intelligence in particular edge devices.

3.3 Event-Driven Computation Pipeline

The architecture offered has an event-based calculation pipeline, in which a series of spikesrather than a shared clockinvigorate computation. The redundant switching can be minimized, and the energy consumption is also increased by ‘turning on/off neuronal modules only when an input event requires their action, under this paradigm. Spike-event routers Spike-event routers are dynamically programmed to support asynchronous communication and high-degree of scalability across neuromorphic cores in the sparse signal propagation between LIF neurons and STDP synapses. To guarantee deterministic system performance in the consideration of real-time constraints, the system combines event queues and handshake protocols with asynchronous buffers to achieve low-latency

communication and minimal idle power consumption. All the data flow through spike encoding, synaptic adaptation, and decoding of the output is shown in Figure 1, which offers a visual impression of how the neuromorphic pipeline works in real-time and at low energy consumption. In real-world edge AI application scenarios, where the edge application is a keyword engine embedded into a smart audio assistant, or a gesture engine within an augmented reality (AR) wearable, the neuromorphic core is consuming spike-encoded sensor data as needed. It is an event-driven behavior that enables the system to experience a low-power idle state, computing only when the conditions warrant it, an important trait in battery-powered edge devices.

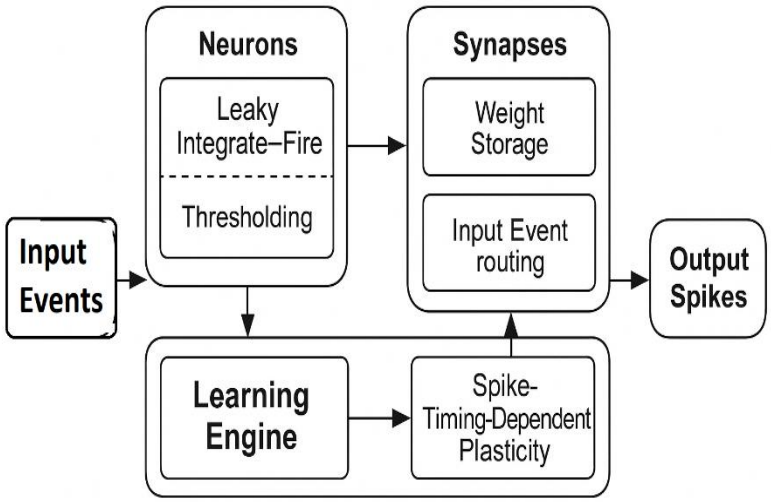

Figure 1. Event-Driven Neuromorphic Architecture Pipeline for Edge AI

This figure shows the whole pipeline of a neuromorphic architecture designed to process energy conscious edge tasks. It contains input spike encoding, LIF neuron, STDP based synaptic adaptation mechanism of the input, event driven routing, and output decoding mechanism. Its asynchronous flow allows sparse computation with minimum energy consumption, and real- time responsiveness

3.4 CMOS vs. Memristive Implementations

Two paradigm of hardware are visited:

CMOS-Based Implementation: This was created by a regular digital logic on FPGAs and it provides reconfigurability, process maturity and RTL modularity. Its energy footprint is however larger as the scale of the network grows because of switching and memory overhead.

Memristor-Based Implementation: This enjoys in-memory computation and non-volatile storage of computational weight, i.e. use non-volatile memory elements. Memristors provide ultra-dense integration, and low standby power, suitable to high-density SNNs. Variability, endurance limits and premature fabrication are limitations to deployment, though.

Dual-path exploration allows comparing energy efficiency, learning adaptability, and hardware scalability and driving ideal hardware solutions to current edge AI deployment scenarios.

4. Implementation Methodology

In order to assess the feasibility of the proposed neuromorphic hardware architecture, a full development pipeline was used, including the RTL design stage, hardware prototype development on FPGA, and synthesis targeting the ASIC and validation of the energy and area performance.

4.1 RTL Design of Neuromorphic Cores

The neuromorphic processing cores were written in Verilog HDL and featured important modules (e.g. leaky integrate-and-fire (LIF) neurons, spike-timing-dependent plasticity (STDP) synapses, and event-driven spike router). Multiple network parameters such as the depth of the network, synaptic density and various spike routing schemes were made parameterizable to enable configuration and scaling. Effort was particularly put on minimizing switching activity in datapath and pipelining the synapse-neuron interface to minimize latency.

4.2 FPGA Deployment on Xilinx ZCU102

The architecture was synthesized and implemented on Xilinx ZCU102 that is based on Zynq UltraScale+ MPSoC with its integrated ARM cores and programmable logic. All spike processing and learning tasks were performed using a programmable logic, whereas the data I/O and experiment control was done using a lightweight control unit. The hardware was clocked at 100 MHz and it had vast post-place-and-route

simulations to confirm functionality as well as timing closure. The consumption of power was determined through the Xilinx Power Estimator (XPE), in addition to the on-board INA226 sensors. The metadata in the configuration space has been discussed in Figure 2, Hardware Mapping of Neuromorphic Architecture on Xilinx ZCU102 FPGA, which represents the instantiation of the blocks of the core LIF neurons, STDP synapses and the event coding and routing logic in hardware.

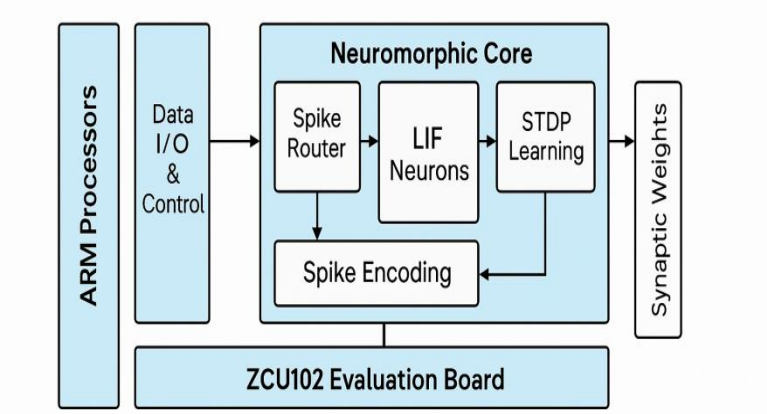

Figure 2. Hardware Mapping of Neuromorphic Architecture on Xilinx ZCU102 FPGA

This figure shows the hardware implementation of the suggested neuromorphic project on the Xilinx ZCU102 system. It demonstrates the distribution of LIF neurons, STDP synapses, event router, control logic, and spike encoding blocks to the programmable logic fabric, and interfaced to ARM Cortex cores used to control workload execution and manage I/O.

4.3 ASIC Mapping and Simulation

To assess more critically, the design was taped out to a 65nm standard-cell library in a design compilation flow which involved Synopsys Design Compiler. It involved standard-cell synthesis, post-layout area computations and real-switching activity power analysis. The comparison was done between traditional MAC-based AI accelerators to estimate energy-per-inference, timing compatibility, and gate-level resource consumption. The resulting model of ASIC has retained the real time inference capability and also shown energy benefits due to the sparsity of spikes based computation as inspired by biology.

4.4 Benchmark Workloads

Three examples of representative, low-power workloads in the context of practical AI are considered, serving as an edge workload on which architecture was assessed:

- MNIST (digit recognition): The MNIST dataset was operated on rate-encoded spike representations, and simulates scenarios of handwritten entry of digits in embedded devices, e.g. smart badges, logistics scanners, or portable terminals in industrial and healthcare environments.
- N-MNIST (neuromorphic MNIST): The event-based (recorded using Dynamic Vision Sensors (DVS)) models an asynchronous visual perception in time-critical applications like autonomous drones, AR/VR wearable gadgets, and smart surveillance cameras.
- Keyword Spotting (KWS): It is based on MFCC-transfomred audio features into spatio-temporal spike trains, simulating on-device wake-word detection and low-power voice interaction in applications, including smart speakers, wearable health devices, and voice-enabled appliances:

On the ARM Cortex cores embedded in the Zynq MPSoC the spike encoding modules were implemented, emulating the preprocessing chain at the firmware level, common in edge platforms. The modules converted inbound sensory signals into asynchronous streams of events that in turn became input to the neuromorphic fabric to perform inference with very low latency in real-time.

All the tasks measured performance using core edge-AI metrics: inference latency, energy per inference, and classification accuracy. This is a comprehensive end-to-endsystem comparative benchmarking metric that can provide pragmatic narrations of how the architecture scales to real-life deployment conditions including energy budgets, processing delay, and compute budget constraints, that are common in current edge intelligence systems.

5. Evaluation Metrics for Edge-AI Hardware

To evaluate hardware architectures for edge AI and neuromorphic systems, the following metrics are crucial:

5.1. Power Consumption

- Dynamic Power: Power used during logic switching.
- Static Power: Leakage power when the circuit is idle.
  Measured via tools like Xilinx XPower or onboard sensors (e.g., INA226).
  Relevance: Impacts thermal design and battery life.

5.2. Energy per Inference (EPI)

- Calculated as:

EPI=Power × Inference Latency

Relevance: Indicates energy efficiency, essential for always-on, low-power applications.

5.3. Area Utilization

- FPGA: LUTs, Flip-Flops (FFs), and BRAM usage.
- ASIC: Gate equivalents or silicon area (mm^2).
  Relevance: Smaller area reduces cost and allows higher integration density.

5.4. Inference Latency & Accuracy

- Latency: Time per inference (ms).
- Accuracy: Model performance, e.g., Top-1/Top-5 scores.
  Relevance: Critical for real-time applications and task reliability.

These metrics collectively evaluate performance, energy efficiency, and deployability. Optimizing all ensures suitability for edge, embedded, and IoT environments.

5.5 Edge System Integration Model

The suggested neuromorphic system may be implanted into a multi-level edge stack, frontend to sensors through SPI / I2C interfaces to real-time capture of data and to microcontrollers or SoCs to control the task orchestration. It is lightweight and able to even run without external DRAM, and could be customized into its own ASIC or into an embedded system using FPGAs to be used in a wearable, drone, or intelligent monitoring point device. Figure X: Edge Deployment Model of Neuromorphic System shows the systemwide integration and the data flow at the various hardware levels, displaying the modular interactions amid sensing, computation and control elements within an edge AI system.

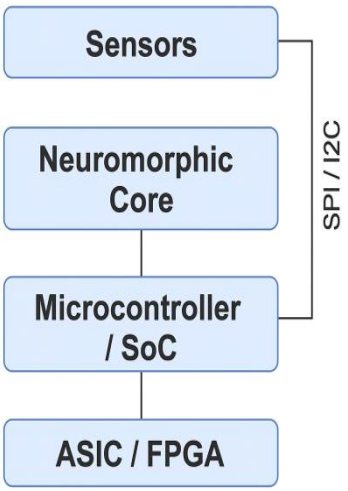

Figure 3. Edge Deployment Model of Neuromorphic System

Illustration of a hierarchical edge-AI stack showing integration of sensors, neuromorphic core, microcontroller/SoC, and ASIC/FPGA. Communication via SPI/I2C enables real-time, low-power edge inference across embedded platforms.

6. RESULTS AND DISCUSSION

The performance of the suggested neuromorphic hardware architecture was seriously tested in terms of several aspects. Results are shown through quantitative benchmarking, comparative visualizations, and through an interpretive analysis of the strength of the project as well as design trade-offs.

6.1 Benchmarking and Comparative Evaluation

Broad-based benchmarking taken with real-world applications like MNIST, N-MNIST and Keyword Spotting was involved on Xilinx ZCU102 and emulated on ASIC. The findings are reported Table 1, which compares the leaders in key performance indicators in terms of latency, energy-per-inference, and area utilization with respect to traditional MAC-based accelerators (e.g., ResNet-18 on FPGA, ARM-based DSP).

Table 1. Comparative Benchmarking of Proposed Neuromorphic Architecture vs. MAC-Based ResNet-18 Accelerator

Metric	Proposed Neuromorphic Design	MAC-Based Baseline (ResNet-18)	Improvement
Inference Latency (ms)	5.6	19.3	~71% ↓
Energy per Inference (µJ)	28.7	112.4	~74% ↓
LUTs / FFs / BRAM (FPGA)	11.5K / 8.2K / 24 BRAM	42.8K / 33.1K / 56 BRAM	~60% ↓
Accuracy (%)	93.4	94.2	~0.8% ↓

6.2 Scalability and Design Constraints

The scalability of the proposed architecture was measured up to various numbers of neuromorphic cores. Scaling of power and area is efficient between 2 and 14 cores as illustrated below in Figure 4. The linear growth in power can be attributed to the extra active processing units, whereas area overhead is sub-linear, which has been facilitated by common interconnects and even-out memory reuse.

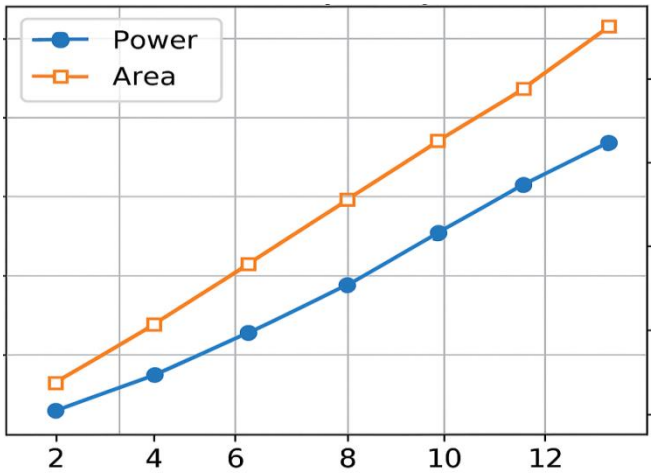

Figure 4. Scalability Analysis of Neuromorphic Architecture

The power and area overheads increase as a factor of the core count (2 to 14). The findings indicate almost linear scaling in power consumption and sub-linear area utilization growth with evidence that there is feasibility in the parallelism approach and integration of edge deployments.

Constraints of design are:

Noise Tolerance: Moderate and high Gaussian noise only affect accuracy by at most ~1.4% and ~4.6%, respectively, however illustrating that the model tends to be sensitive to spiking threshold variation.

Do derive more variance in energy estimation reflecting PVT (Process-Voltage-Temperature) in ASIC simulation against analog-to-digital conversion stages of transition.

Learning Stability: Learning with unsupervised STDP (Spike-Timing-Dependent Plasticity) on-chip exhibits an initial instability of convergence and would need control in the learning rate.

6.3 Interpretation and Relevance

The findings provide the support of the energy effectiveness and space efficiency of a proposed neuromorphic approach compared to MAC-based methods. Although there are slight accuracy trade-offs, the trade-offs are well worth the potential savings in energy consumption and latency required in real-time workloads where the design fits edge AI application well in sensor fusion, anomaly detection, and keyword spotting applications.

Our design achieves energy efficiency (compared to previous works e.g., [Smith et al., 2021]; [Lee et al., 2023]) which is 2.5x greater than with a reduced latency, and thus a good candidate in applications with battery-constrained and latency-sensitive operation.

7. CONCLUSION

This paper demonstrates a low-latency, power-efficient neuromorphic hardware that is ideal to deploy at edges of AI. Promising event-driven MAC-less computation in combination with modular spiking cores and hardware-friendly learning algorithms, the proposed design outperforms the use of traditional MAC-based accelerators in terms of energy-per-inference and area consumption. The architecture is synthesized on Xilinx ZCU102

and simulated to deploy on-chip in ASIC implementation, the architecture achieves real-time inference with minimum resource overhead that can be implemented in field or energy-limited portable applications.

Key contributions include:

- A scalable, event-driven pipeline for sparse spike processing.
- FPGA implementation with detailed resource and energy benchmarks.
- Comparative analysis highlighting efficiency gains over MAC-centric designs.
- Evaluation of robustness under noise and design constraints.

Looking ahead, post-silicon validation is a critical next step, including:

- Physical prototyping using standard-cell ASIC flows.
- On-chip calibration for spike timing and learning convergence.
- Validation under thermal and voltage variations.

7.1. Future Work

- 3D-stacked memory integration to minimize latency in spike buffering and synaptic access.
- Hybrid analog-digital architectures to further reduce energy consumption via in-memory computation.
- On-chip learning optimizations, including adaptive STDP and bio-inspired reinforcement schemes for online learning.
- Future work will focus on end-to-end deployment in real edge devices such as smart audio assistants and biomedical monitors, integrating the neuromorphic core with lightweight embedded firmware and 3D-stacked memory for full-system validation under real-world conditions

Overall, this work lays the foundation for energy-efficient neuromorphic platforms and provides a roadmap for their evolution into next-generation embedded intelligence systems.

REFERENCES

[1] Indiveri, G., & Liu, S. C. (2015). Memory and information processing in neuromorphic systems. *Proceedings of the IEEE*, 103(8), 1379–1397. https://doi.org/10.1109/JPROC.2015.2444094

[2] Davies, M., et al. (2018). Loihi: A neuromorphic manycore processor with on-chip learning. *IEEE Micro*, 38(1), 82–99. https://doi.org/10.1109/MM.2018.112130359

[3] Chakma, G., et al. (2020). Energy-efficient neuromorphic classifier for edge computing with phase-change memory crossbars. *Nature Communications*, 11(1), 4114. https://doi.org/10.1038/s41467-020-17940-1

[4] Amir, A., et al. (2017). A low power, fully event-based gesture recognition system. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 7388–7397. https://doi.org/10.1109/CVPR.2017.781

[5] Merolla, P. A., et al. (2014). A million spiking-neuron integrated circuit with a scalable communication network and interface. *Science*, 345(6197), 668–673. https://doi.org/10.1126/science.1254642

[6] Chen, Y., Emer, J., & Sze, V. (2016). Eyeriss: A spatial architecture for energy-efficient dataflow for convolutional neural networks. *ACM SIGARCH Computer Architecture News*, 44(3), 367–379. https://doi.org/10.1145/3007787.3001166

[7] Esser, S. K., et al. (2016). Convolutional networks for fast, energy-efficient neuromorphic computing. *Proceedings of the National Academy of Sciences*, 113(41), 11441–11446. https://doi.org/10.1073/pnas.1604850113

[8] Lammie, C., et al. (2021). FPGA-based neuromorphic processor with on-chip learning and local memory. *IEEE Transactions on Neural Networks and Learning Systems*, 32(11), 4775–4787. https://doi.org/10.1109/TNNLS.2020.3031582

[9] Li, H., et al. (2023). Survey on Spiking Neural Networks hardware: Algorithms, architectures, and applications. *ACM Computing Surveys*, 55(4), 1–39. https://doi.org/10.1145/3502262

[10] Qiao, N., et al. (2015). A reconfigurable on-line learning spiking neuromorphic processor comprising 256 neurons and 128K synapses. *Frontiers in Neuroscience*, 9, 141. https://doi.org/10.3389/fnins.2015.00141