SENNA: An Ultra-Low Latency, Time-Deterministic Field-Programmable-Spiking-Neuron-Array (FPSNA)

Ferdinand Pscheidl[1], Elmar Herzer[2], Andreas Stoll[2], Leonidas Richter[2], Claudia Schuhmann[2], Stefan Köhler[2], Sebastian Ruping[1], Matthias Kampa[3], Ralf Brederlow[4], Johannes Leugering[5]
Email: ferdinand.pscheidl@emft.fraunhofer.de, elmar.herzer@iis.fraunhofer.de, jleugering@ucsd.edu
[1]Fraunhofer EMFT, Munich, Germany, [2]Fraunhofer IIS, Erlangen, Germany, [3]now with ZHAW, Zürich, Switzerland, [4]Technische Universität München, Munich, Germany, [5]now with UCSD, San Diego, CA

***Abstract*— SENNA is a first-of-its-kind mixed-signal circuit-switched field-programmable accelerator for Spiking Neural Networks (SNNs). We demonstrate a 1024-neuron and 25.6k synapse proof-of-concept chip in 22nm FDSOI that achieves 687 fJ energy per synaptic operation (E/SOP) at 25 MHz and 23 mW. The IC allows a very flexible configuration of parallel processed SNNs. Due to its low power draw, low spike-in-spike-out (SISO) latency of 40 ns and time-deterministic response, this design is particularly promising for real-time edge applications.**

Keywords—neuromorphic, SNN, real-time, low latency

I. Introduction

Neurons in biological neural networks encode information in the relative timing of spikes. The field of neuromorphic engineering has long proposed spike-based processing as a blueprint for power efficient, massively parallel, mixed-signal accelerators for artificial intelligence (AI) [1]. Current state-of-the-art neuromorphic processors use time-multiplexing, packet-switched network-on-chip (NoC) architectures and interfaces based on the address-event-representation (AER) protocol [2-6] to route spike messages. This introduces data-dependent delays and risks congestion, thus losing critical information encoded in the relative spike-timing. As a result, most neuromorphic processors decouple simulated and wall-clock time to ensure algorithmic correctness in exchange of latency determinism.

A more "brain-like" solution with hard-wired synaptic connections can achieve state-of-the-art E/SOP with deterministic delays, but at the cost of flexibility and area [7]. Programmable switchboxes as used in FPGAs provide a more versatile alternative. Circuit-switched, fully-parallel, point-to-multi-point connections between neurons reduce latency and power by obviating address headers and routing logic, and by exploiting sparse, static, small-world connectivity [8-9]. Crucially, this approach preserves deterministic timing, supports multicasting natively and is field-programmable. Combined with an array of parallel, time-continuous, mixed-signal spiking neurons [7], we call this a Field-Programmable Spiking Neuron Array (FPSNA). Here we present the novel SENNA architecture and a proof-of-concept implementation in 22nm FDSOI, designed for small- to mid-scale real-time edge-applications and our mixed-signal and mixed-timing design flow that leverages standard digital design automation tools.

II. SENNA

The *SENNA* architecture (Fig. 1) combines a FPSNA for SNN computation with a direct spike interface, on-chip spike generators (Fig. 2) and recorders, a serial interface for configuration, a global controller and a clock generator in a digital-on-top, mixed-signal and mixed-timing design flow. The FPSNA comprises an asynchronous, digital, circuit-switched mesh for programmable point-to-multi-point connections between neurons with one mixed-signal spike processing node (SPNs) and associated switchbox (SWB) per grid position. Each SPN contains one neuron with clocked spiking output and synaptic weights. Synaptic connections of a SNN are mapped onto SENNA by configuring both synaptic weights and SWBs (Fig. 3). Unlike packet-switched digital designs, our mixed-signal SPNs can receive and process input spikes time-continuously and fully parallel. In addition to the high-speed, direct spike interface, which can process up to 25 million external spikes/second per channel, spike patterns can also be generated and recorded on-chip via pattern generators, pattern players, recorders and counters, which can be programmed and read out via a serial interface.

The clock generator creates three clock domains from an external reference clock. The first clock runs controller and on-chip memory, the second clock controls spike generators and recorders, and the third clock the spike generation in SPNs. Clocks two and three are divided from clock one, are balanced to each other and have a configurable duty cycle, that controls the spike widths. The phase of output spikes of the mesh is synchronized to clock two with a custom digital cell for phase synchronization such that spikes can be synchronously recorded. Thereby the different clock domains enable running control, neurons and spike inputs and outputs on different time scales.

Fig. 4 shows the life cycle of a spike and the clock domain crossings it experiences. Our mixed-timing digital design flow

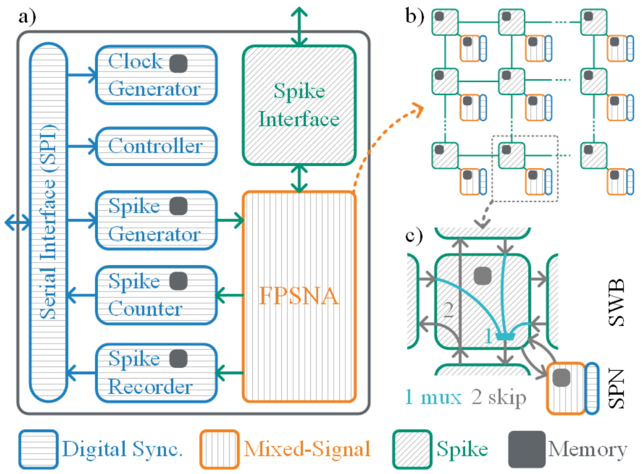

Fig. 1. a) SENNA architecture top-level. b) Zoom-in on the FPSNA with SWB and SPN per grid position. c) Zoom-in on one grid position displaying examples of the two possible connection types: Configurable mux connection (1) and fixed skip connection (2).

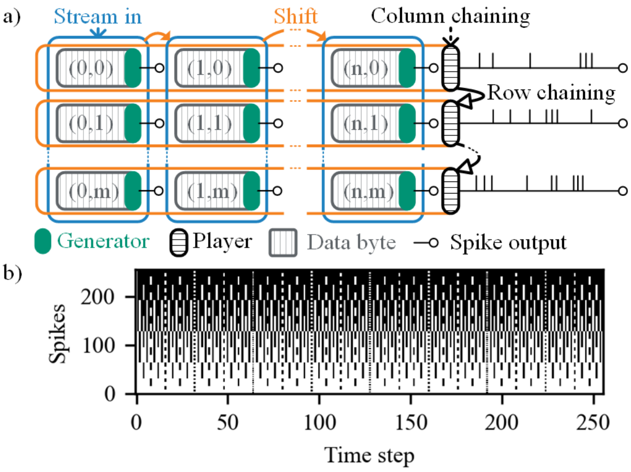

Fig. 2. a) n×m in-memory spike generators are programmed by stream-in and shift. One pulse density pattern generator per data byte can generate patterns as shown in b). Column chaining allows bitwise pattern playing. Row chaining enables longer patterns.

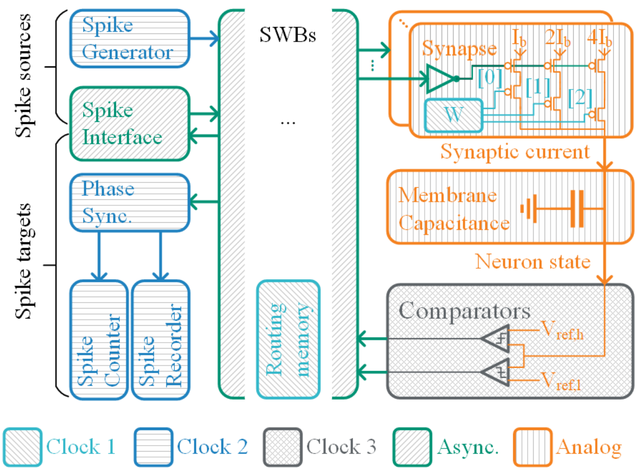

Fig. 4. Possible life cycles and timing domain crossings of a spike. The parallel neuron synapses can process asynchronous spike inputs time continuously. Phase synchronizers enable synchronous spike outputs.

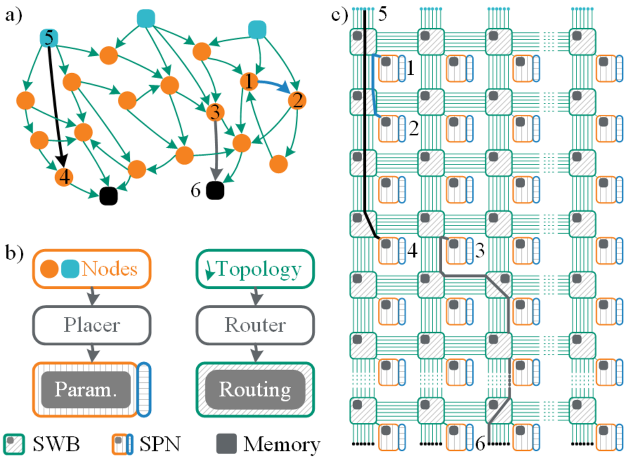

Fig. 3. Programming of FPSNA. a) Example RSNN. b) Placing nodes onto physical inputs, SPNs and outputs and routing the SNN topology through SWBs. Parameters, e.g. weights, are stored in SPNs, topology in SWBs. c) Example of routes in FPSNA.

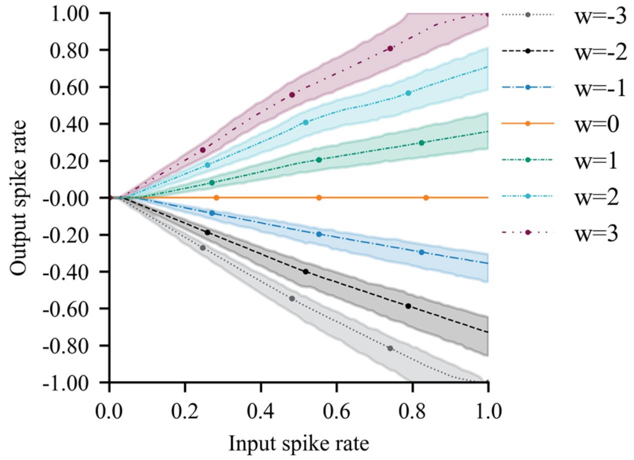

Fig. 5. Measured neuron gain for different weight (w) configurations. Halos show range of results for different synapses of same SPN. Only one weight is active at a time. Weight scaling is set to 1 and feedback to 3. Negative output spike rates correspond to nspikes, positive rates to pspikes.

uses synchronous timing constraints to ensure functionality of logic and provide accurate timing references for spike generation. Once a spike is generated and enters the circuit-switched mesh, which is completely digital unlike [9], it is considered asynchronous. Its implementation can be integrated into a timing-driven, digital place and route flow, resulting in a scalable design flow for this mixed-signal architecture, even for strongly heterogeneous SWBs.

A. *SENNA Prelude*

The proof-of-concept chip *SENNA Prelude* implements a 32x32 FPSNA in 22nm FDSOI technology. Each SPN is a leaky-integrate-and-fire (LIF) neuron [7] with 25 3-bit switched current-mirror synaptic weights, a charge-based neuron membrane state and two complementary outputs, one for a positive threshold detection (pspike) and one for a negative threshold detection (nspike) [10]. The threshold detection is implemented via clocked comparators running at up to 25 MHz. A local, configurable bias current generator per neuron enables weight scaling for synaptic weights. Each neuron has configurable switched capacitance leakage and its own local spike generator. Unlike [10], we use negative feedback instead of reset after each threshold detection to improve linearity. We use custom distributed, serial-in and parallel-out SRAM holding two dynamically selectable configurations for SPNs, SWBs and the spike generators. A total of 34x64 in-memory spike generators can either generate regular pulse patterns with a density given by an 8-bit integer or play precise patterns of variable length from 8 to 4352 stored as a bit-vector. Each SWB comprises 100 inputs and 100 outputs per neighbor. Up to 109 2x8-bit SRAM cells configure switchable (mux) connections using 2-, 4-, 8- and 16-input multiplexers; the remaining connections are fixed (skip). 128 FPSNA outputs can be observed by 8-bit spike counters. The timing of up to 82k spikes from 108 of the outputs can be recorded by SRAM macro-based spike recorders using phase synchronization circuits and a reference clock. The netlist of the FPSNA was software-generated based on a set of connection rules. We used a digital-on-top design flow for physical integration and STA, and verified the mixed-signal SPNs, developed in an analog design flow, against a behavioral SystemVerilog model.

B. *Executing SNNs on the FPSNA*

Running a SNN on *SENNA* requires similar steps as running RTL designs on FPGAs. First, a suitable SNN must be found, then it can be placed and routed on the FPSNA. Finally, a bitstream file is generated to program *SENNA*.

TABLE I. Comparison with state-of-the-art neuromorphic processors.

	NCE, 2024 [2]	MICRO, 2018 [5]	JSSCC, 2019 [3]	BioCAS, 2024 [7]	PIEEE, 2019 [4]	This work
Implementation	Mixed-Signal	Digital (async)	Digital	Mixed-Signal	Mixed-Signal	Mixed-Signal
Technology	180nm	14nm FinFET	10nm FinFET	22nm FDSOI	28nm FDSOI	22nm FDSOI
Area, mm²	69.94	60	1.72	0.014	0.65	10.48
Neurons	1024	131k	4096	21	4096	1024
Synapses	64k	14.6M – 128M	1M	2310	65k	25.6k
Spike communication	NoC	NoC	NoC	Fixed	NoC	FPSNA
State update	Fully-parallel	Time-multiplexed	Fully-parallel	Fully-parallel	Time-multiplexed	Fully-parallel
Max. Spike rate	Async < 7kHz	Async < 5Mhz	Conf. < 10MHz	50 MHz	n.a.	25 MHz[1]
Versatility	Med.	High	High	Low	Med.	**Med.**
Time det.	No	No	No	Yes	No	**Yes**
Min. SISO	n.a.	n.a.	n.a.	60ns	n.a.	**40ns**
E/SOP	150 pJ	23.6 pJ	3.8 pJ	86.9 fJ	380 fJ	**687 fJ**

[1]Designed for and simulated at 100 MHz, measured successfully so far only at 25 MHz.

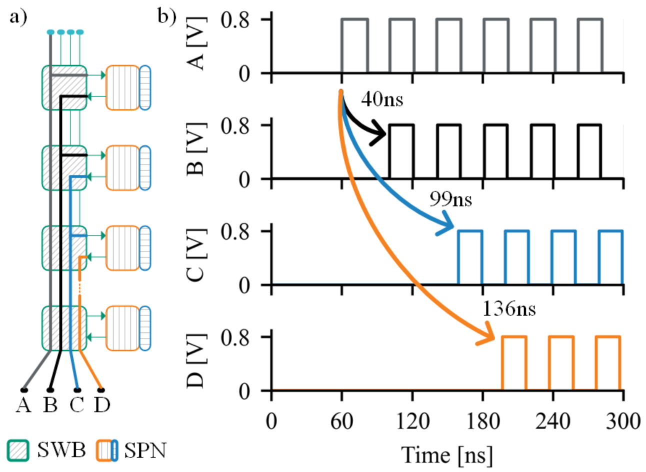

Fig. 6. Latency measurement between four spike outputs A, B, C and D. A is a forwarded spike input. B, C and D are each an output of neurons connected in a chain to the input.

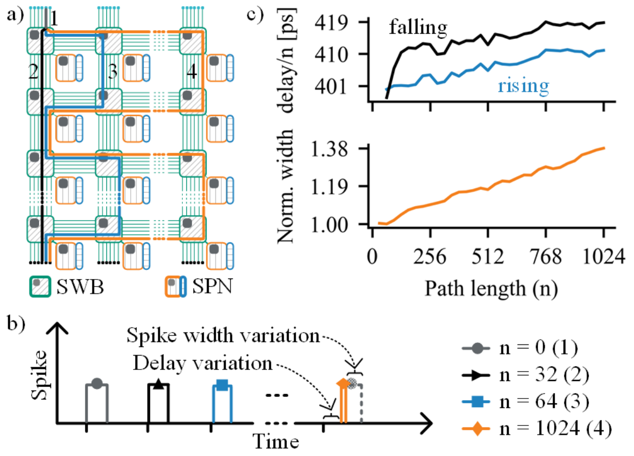

Fig. 7. a) Routing configurations for path lengths of n € {32, 64, 96, ⋯, 1024}. b) Visualization of delay and spike width variation due to asymmetric fall and rise times. c) Measured average falling and rising delay per hop and normalized spike width for different path lengths.

Since shorter connections occupy fewer resources in SWBs, *SENNA* is best suited for small-world SNNs as found in biological brains. Moreover, the homogeneous and symmetric SWB design used in *SENNA Prelude* achieves the best utilization of hardware resources in networks with recurrent connections.

The synchronous spike generation in SPNs and dedicated spike generators make it possible to model and train SNNs running on *SENNA* with existing frameworks that use time-stepped simulation, without the need to introduce artificial simulation time-steps. The deterministic latency of the internal NoC removes another potential source of timing mismatch between simulation and hardware. Maximum delay limits can be considered during the place-and-route phase. Other sources of error are voltage and temperature dependent delays, variations in synaptic weights and neuron parameters, like threshold and weight scaling, or asymmetric signal rise and fall times in SWBs that cause spike width variation. All could be mitigated in part by hardware-aware or hardware-in-the-loop training.

The spike-in-spike-out (SISO) latency of a SNN on *SENNA* via the direct spike interface defines its minimum response latency to an input spike event. The latency is deterministic and its lower bound is limited by the minimum number of SPNs and SWBs on the critical path between input and output and the maximum spike rate of the neurons. In contrast to neuromorphic processors with only serial interfaces and memory-based, on-chip spike generation [2-6], *SENNA* can thus seamlessly integrate into fully spike-based real-time neuromorphic systems.

III. Measurement Results and Comparison

The *SENNA Prelude* chip was fabricated in 22nm FDSOI technology and wire-bonded in a JLCC68 package. Fig. 9 shows a die photo annotated with layout zoom-ins. We performed measurements at a maximum spike rate of 25 MHz.

In Fig. 5 the neuron gain of a SPN, shown as the output spike rate over the input spike rate, is characterized for different synaptic weight values. We fed synapses with local spike generators and counted the output spikes with spike counters. The output spike rates for different weight values are clearly separated. For different synapses of the same SPN the neuron gain varies by up to 48% for the same weight value.

We measured the SISO from spikes of a single generator through a chain of three SPNs by observing the generated spikes and outputs of all three neurons with a logic analyser via the spike interface (Fig. 6). We derived the SISO for each neuron as the time difference between its first output spike and the generated spike that caused it. Compared to [7], where a 3-layer SNN is implemented, we achieve a 1.5x smaller SISO of 40 ns due to the increased versatility. Other neuromorphic processors [2-6] do not report their SISO, which can’t be deterministic but we expect it to be orders of magnitude larger due to spike rates, memory accesses or serial communication.

To characterize the average delay per switchbox and the spike width variation we measured the delay between an external input spike and the spike taking paths of different lengths n € {32, 64, 96, ⋯, 1024} (Fig. 7). The average rising and falling delays per switchbox vary by up to 2.5% between

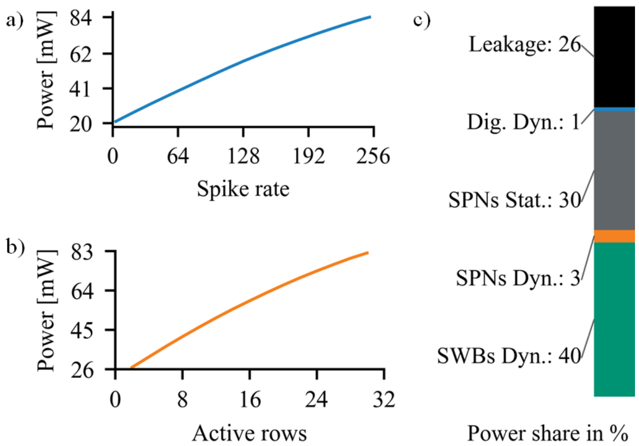

Fig. 8. Power consumption over a) spike rate and b) active SWB rows. c) Average power consumption break-down of LSM. Dig. Dyn. includes all digital modules except SWBs. Dyn. means dynamic and Stat. means static.

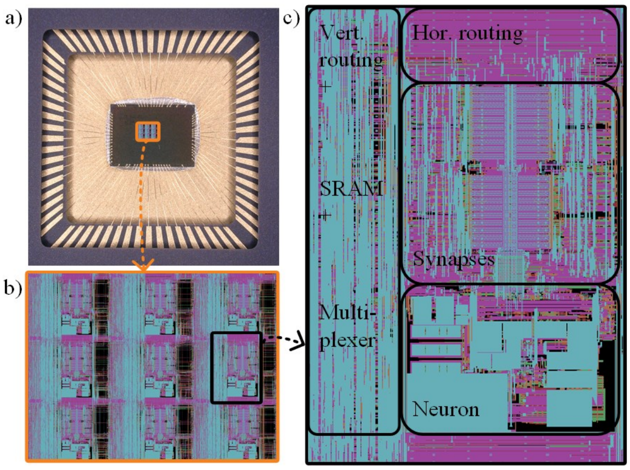

Fig. 9. a) Die photo of SENNA Prelude bonded in JLCC68 package with zoom-ins on b) FPSNA grid and c) SWB vertical routing with SRAM and multiplexers, SWB horizontal routing, SPN synapses and SPN neuron.

the measurements and are ca. 410 ps. At 25 MHz this would mean that a spike of 20 ns can travel a maximum of 48 SWBs without overlapping into the next time step. Reducing the spike rate relaxes the delay constraint. The spike width increases by approximately 8 ps or 0.02% of a clock cycle per switchbox. Increasing the spike width reduces the relative error per switchbox, but increases the E/SOP.

Fig. 8 shows how we measured the FPSNA's routing power performance. First, we iterated over the spike generator spike rates, while configuring all SWBs to forward the generated spikes. Then we configured the generators with their maximum spike rate of 256 and iterated over the number of forwarding SWB rows. We measured the average power consumption over 256 time-steps. The power increased by 0.25 mW per spike rate and by 1.8 mW per active row. From the former we derived an average of 3.1 fJ per spike per SWB connection, which is more than 500 times smaller than the best reported 1-hop energy of 1.6pJ in [9] at 130 nm.

To evaluate the use-case specific E/SOP, we designed a Liquid State Machine (LSM) using all 1024 neurons of SENNA Prelude, connected by 9174 synaptic connections. We determined its linear separation property, the number of linearly separable input patterns, to be around 700, and estimate a VC-dimension, indicating the generalization capability, to be around 100, following [11]. We mapped it to hardware using a naive direct placement and a simple version of the PathFinder [12] routing algorithm. We measured its average power consumption over 256 time-steps to be 23.1 mW. Fig. 8 shows a break-down. With a total of 345k synaptic operations, measured in a corresponding software simulation, we derived an E/SOP of 687 fJ, following [7].

Even though the energy per synaptic operation is higher than [7], other than there our system allows a very flexible and programmable configuration of the spiking neural networks. Besides [7] only SENNA has deterministic latency.

IV. Conclusion

We have demonstrated, to our knowledge, the first energy-efficient, field-programmable SNN accelerator with deterministic timing and more than 640 TSOPs/s.

Although some digital solutions [5] can achieve a higher neuron and synapse density e.g. through time-multiplexing and dense off-chip memory, the low and deterministic latency makes FPSNAs a promising and versatile platform for small world SNNs, especially in neuromorphic event-based sensor and actuator systems. A time-continuous interaction with analog signals via spike-encoding and -decoding circuits could make *SENNA* a programmable alternative to analog ASICs for e.g. high-speed control problems.

References

[1] C. Frenkel, D. Bol and G. Indiveri, "Bottom-Up and Top-Down Approaches for the Design of Neuromorphic Processing Systems: Tradeoffs and Synergies Between Natural and Artificial Intelligence," in Proceedings of the IEEE, vol. 111, no. 6, pp. 623-652, June 2023.

[2] O. Richter, C. Wu, A. Whatley, G. Köstinger, C. Nielsen, N. Qiao and, G. Indiveri,, "DYNAP-SE2: a scalable multi-core dynamic neuromorphic asynchronous spiking neural network processor" in Neuromorphic Computing and Engineering, vol. 4., no. 1, Jan. 2024.

[3] G. K. Chen, R. Kumar, H. E. Sumbul, P. C. Knag and R. K. Krishnamurthy, "A 4096-Neuron 1M-Synapse 3.8-pJ/SOP Spiking Neural Network With On-Chip STDP Learning and Sparse Weights in 10-nm FinFET CMOS," in IEEE Journal of Solid-State Circuits, vol. 54, no. 4, pp. 992-1002, April 2019.

[4] A. Neckar et al., "Braindrop: A Mixed-Signal Neuromorphic Architecture With a Dynamical Systems-Based Programming Model," in Proceedings of the IEEE, vol. 107, no. 1, pp. 144-164, Jan. 2019

[5] M. Davies et al., "Loihi: A Neuromorphic Manycore Processor with On-Chip Learning", in IEEE Micro, vol. 38, no. 1, pp. 82-99, Feb. 2018.

[6] M. Yao et al., „Spike-based dynamic computing with asynchronous sensing-computing neuromorphic chip", in Nature Communications, vol. 15, May 2024.

[7] M. Ochs, M. Dietl and R. Brederlow, "A Circuit Concept for Energy-Efficient Spiking Neural Network Systems with a FOM of 86.9fJ/SOP," 2024 IEEE Biomedical Circuits and Systems Conference (BioCAS), Xi'an, China, pp. 1-5, Oct. 2024.

[8] M. Rathore and G. S. Rose, "SpiCS-Net: Circuit Switched Network on Chip for Area-Efficient Spiking Recurrent Neural Networks", 2024 37th International Conference on VLSI Design and 2024 23rd International Conference on Embedded Systems (VLSID), Kolkata, India, pp. 204-209, Jan. 2024.

[9] T. Dalgaty et al., "Mosaic: in-memory computing and routing for small-world spike-based neuromorphic systems", in Nature Communications, vol. 15, Jan. 2024.

[10] R. Serrano-Gotarredona, T. Serrano-Gotarredona, A. Acosta-Jimenez and B. Linares-Barranco, "A Neuromorphic Cortical-Layer Microchip for Spike-Based Event Processing Vision Systems," in IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 53, no. 12, pp. 2548-2566, Dec. 2006

[11] R. Legenstein and W. Maass, „Edge of chaos and prediction of computational performance for neural circuit models", in Neural Networks, vol. 20, no. 3, pp. 323-334, April 2007.

[12] L. McMurchie and C. Ebeling, "PathFinder: A Negotiation-Based Performance-Driven Router for FPGAs," Third International ACM Symposium on Field-Programmable Gate Arrays, Napa Valley, CA, USA, pp. 111-117, Feb. 1995.