1.7pJ/SOP Neuromorphic Processor with Integrated Partial Sum Routers for In-Network Computing

B. Wang[1], M. M. Wong[2], D. Li[1,2], Y. S. Chong[2], J. Zhou[3], W. F. Wong[3], L. Peh[3],
A. Mani[2], M. Upadhyay[3], A. Balaji[3], and A. T. Do[2]
[1]Singapore University of Technology and Design (SUTD),
[2]Institute of Microelectronics (IME), Agency for Science, Technology and Research,
[3]Department of Computer Science, National University of Singapore (NUS), Singapore.
bo_wang@sutd.edu.sg[1]

***Abstract*—Conventional neuromorphic accelerators primarily leverage split-merge method to accommodate a neural network that is beyond a single core's size, leading to possible accuracy loss, extra core usage and significant power and energy overhead. This work presents an energy-efficient, reconfigurable neuromorphic processor to address the problem by (i) a partial sum router circuitry that enables in-network computing to remove the need of extra merge cores; (ii) software-defined Networks-on-Chip that eliminates the power-hungry routing compute and (iii) fine-grained power gating and clock gating technique for power reduction. Our test chip achieves lossless mapping as the algorithm and an energy efficiency of 1.7pJ/SOP at 0.5V, 19% lower than state-of-the-art result.**

***Index Terms*—Neuromorphic Processor, Energy Efficient, Network on Chip, In-network Computing.**

I. Introduction

Recent Spiking Neural Network (SNN) processors have successfully demonstrated impressive performance and energy efficiency in digit recognition [1], object detection [2], robotic control [3] and event-based perception [4], etc. However, typical SNN architectures leverage on neuron cores for synaptic weight accumulation while deploying Networks-on-Chip (NoCs, i.e. buffers, routers and wire links) for spike delivery across the cores [5], [6], [7]. This can lead to information loss due to spike quantization error that arises from the size constraint of a physical neuron core. As Fig. 1(a) depicts, a post-synaptic neuron receives a number of synaptic weights from the pre-synaptic neurons. A single neuron core can not accommodate all the synaptic inputs if they exceed the fan-in limit of the core. Hence a many-fan-in layer (e.g., a fully connected layer) has to be mapped onto multiple cores whereas each core individually fires a spike based on a partial sum instead of a total sum. The intermediate spikes are subsequently collected using additional merge cores such that an approximated total sum is formed with retrained weights and thresholds, namely split-and-merge mapping [1], [2]. However, the exact partial sum values have lost after the quantization in the split cores, rendering the total sum discrepant from the software-computed weighted sum and eventually degrading inference accuracy. Moreover, the split-and-merge scheme incurs considerable area and power overheads due to the additional merge cores, suppressing the energy efficiency advantage of utilizing SNN.

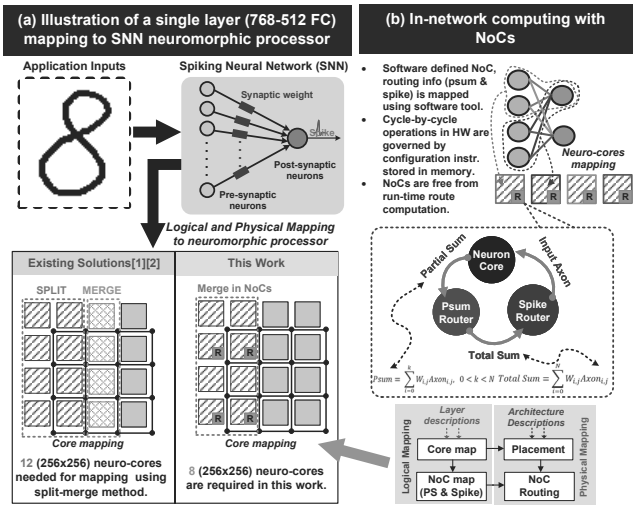


Fig. 1. (a) Low-overhead core mapping for proposed neuromorphic processor; (b) Proposed In-network computing technique facilitated by partial-sum and spike NoCs.

To overcome the issue, we propose In-Network Computing (INC) and demonstrate it with Shenjing architecture [8] via compute-enabled NoCs to achieve lossless, flexible and energy-efficient SNN operations (Fig. 1(b)). The proposed technique is able to (1) prevent the accuracy loss due to the neuron core quantization by enforcing partial sums to aggregate within the NoCs till a full total sum is formulated and (2) allow efficient core mapping with less physical cores by eliminating the need of merge cores. Moreover, our NoCs circuitry offloads the routing computation and the flow control from the neuromorphic processor to the software. To further improve area efficiency, our processor reuses a single partial-sum/spike router for all the fan-out neurons in a core via time multiplexing whereas Shenjing [8] adopts dedicated routers for

The work was partially supported by NRF, Singapore (NRF-CRP23-2019-0003) and SUTD, Singapore (SRT3IS20162, SGP-AIRS1841, SUTD-ZJU(VP)201808)

every neuron. Enhanced with power gating and clock gating, the chip in 40nm CMOS technology consumes merely 0.5 mW power and 1.7pJ energy per synaptic operation with lossless mapping from the algorithm onto the hardware.

II. Lossless Mapping with In-Network Computing

Fig. 2(a) details the mapping of an illustrative $768\times512\times10$ SNN onto the processor for MNIST digit inference. Layer 1 (i.e., L1 - 768×512) is mapped to the first 8 cores (i.e., red cores) of the 4×3 array while Layer 2 (i.e., L2 - 512×10) is mapped to core (0, 2) and (1, 2) (i.e., yellow cores). Considering L1, instead of quantizing the partial sum within each core, the partial sum router integrates an adder for accumulation along the data propagation till it arrives at core (0, 0) and core (0, 1) where the final spikes are sent to L2. Eventually, the dataflow forms an adder tree (Fig. 2(b)), starting from the 8 cores and aggregating at the root node core (0, 2) where the classification is made. In other words, the dataflow is accelerated along propagation as the NoCs possess computing capability to aggregate it, which is the essence of In-Network Computing. This allows direct mapping of a logical core on the SNN hardware without retraining to recover the classification accuracy loss due to the binary decision of the spiking function within each core. Note that we still require multiple cores to map a large layer, but the number of cores is 33% lesser than the split-and-merge scheme [1], [2], thanks to the partial sum router. Fig. 2(c) depicts the circuit diagram of a Neural Process Unit (NPU or Neuro-core), including a neuron core, a partial sum router and a spike router. The neuron core is implemented with SRAM banks that transfer the multiplication of spike input and synaptic weight into memory read. The local partial sum from each neuron can be injected to the partial sum NoCs via output ports (N/S/E/W) or added with the incoming partial sums for multiple times in the router. The router addresses the routing requirements of the 256 neurons one by one via time multiplexing. All the operations are orchestrated by an FSM controller. The details of the routing operation are elaborated in Section III.

III. Proposed Neuromorphic Processor

A. Low Power Circuit Techniques

Low power circuit techniques are incorporated in the SNN chip to further reduce its power and improve the energy efficiency (Fig. 3). First, the neuron core is divided into 2×2 sub-cores, each implemented with 128×128 SRAM banks such that the two vertical cores can be operated in parallel to improve the throughput and reduce latency (Fig. 3(a)). This partitioning of SRAM banks at the sub-core level allows fine-grained power gating in the synaptic memory when a smaller number of neurons are needed. In the scenario where only one sub-core is in operation, we can suppress the dynamic and the leakage power values of an NPU by 44% and 50%, respectively by enabling power gating on the remaining sub-cores (Fig. 3(b)). As synaptic activities can be determined offline by using our in-house compilation flow, sub-core power gating control bits are set and programmed into the chip,

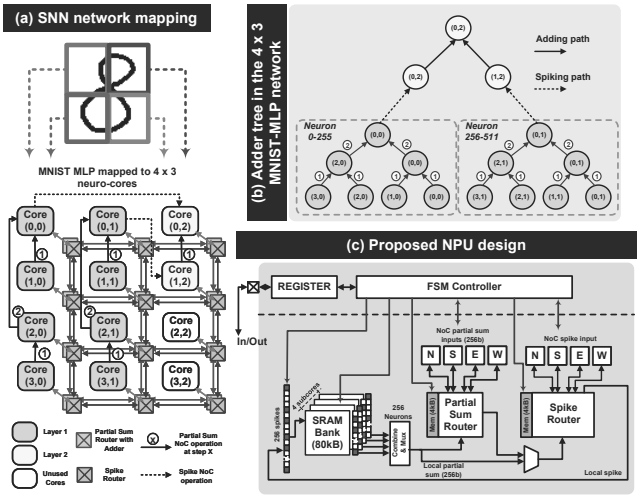
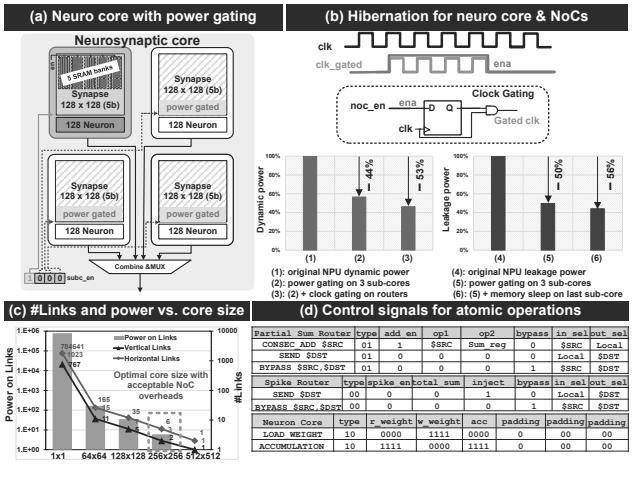


Fig. 2. (a) Illustration of MNIST-MLP mapping with the proposed processor; (b) Adder tree for partial sum addition to eliminate accuracy loss; (c) Detailed architecture of the NPU design.

depending on the network requirement. Second, at the core level, memory sleep and clock gating are applied to the sub-cores and the router circuits, respectively. In conjunction with the power gating technique, the dynamic and the leakage power per NPU can be eventually reduced by 53% and 56%, respectively (Fig. 3(b)). Third, the number of NoC links and the power associated with the core dimension are investigated to determine the optimal core size (i.e., 256×256) where the power overhead and the SRAM utilization is balanced (Fig. 3(c)). Finally, atomic-operation-level configuration is supported by the processor via control signals from our toolchain. This facilitates efficient core mapping and effective energy calculation by decomposing a complex task into a population of atomic operations. An example is illustrated in Fig. 3(d).

B. NoC Routers and NPU

The implementation of NoC routers is illustrated in Fig. 4. In the partial sum router, the output of the adder is registered and fed back to the accumulation data path for consecutive additions. The receiving input, if not needed locally, will bypass adjacent cores instantly. Similarly, the spike router registers bypass the incoming spikes as a crossbar switch between the input and the output ports. The routing operations are defined by the instructions that are computed offline and preloaded in a tiny register file and a Look-Up Table (LUT), depending on the mapping result. All the 256 neurons in the core, connect to one partial-sum router and one spike router, respectively in a time multiplexing manner. The interconnected routers form a dedicated link that guarantees the partial sum or spike from the current neuron has been delivered to the destination before switching to the next neuron. To offset the multiplexing-induced delay, the router works at a $256\times$ higher frequency compared to the neuron circuit, ensuring to serve all the neurons in one clock cycle from the neuron's perspective.



Partial Sum Router	type	add_en	op1	op2	bypass	in_sel	out_sel
CONSEC_ADD $SRC	01	1	$SRC	Sum_reg	0	$SRC	Local
SEND $DST	01	0	0	0	0	Local	$DST
BYPASS $SRC,$DST	01	0	0	0	1	$SRC	$DST

Spike Router	type	spike_en	total_sum	inject	bypass	in_sel	out_sel
SEND $DST	00	0	0	1	0	Local	$DST
BYPASS $SRC,$DST	00	0	0	0	1	$SRC	$DST

Neuron Core	type	r_weight	w_weight	acc	padding	padding	padding
LOAD WEIGHT	10	0000	1111	0000	0	00	00
ACCUMULATION	10	1111	0000	1111	0	00	00

Fig. 3. (a) Sub-cores can be de-activated individually in a neuron core by power gating; (b) Core and routers hibernate when not in use for power saving; (c) Optimal core size vs. power w.r.t. MNIST L1 layer; (d) Atomic operations supported in the instruction set to control neuron core and NoCs.

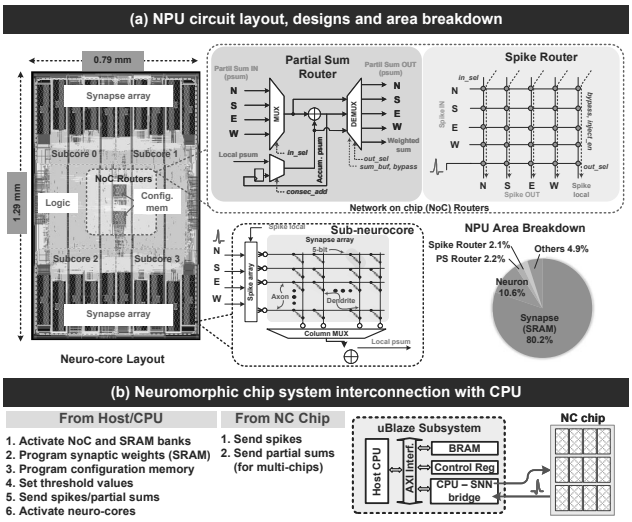
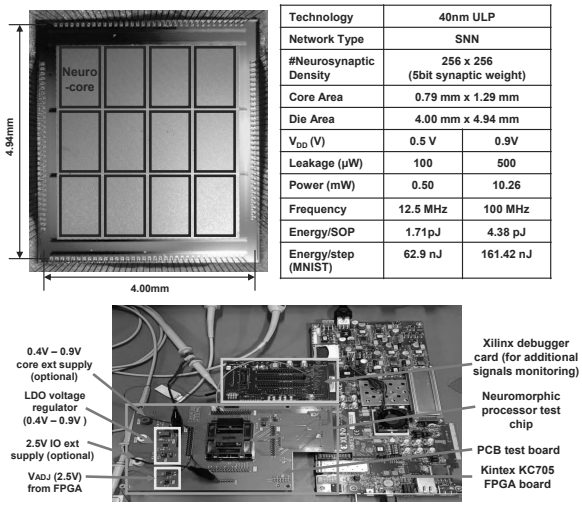


Fig. 4. (a) NPU layout and circuit designs for the sub-neuron core and NoCs with area breakdown; (b) Microblaze (FPGA) is used to connect the Neuromorphic Chip (NC) and the CPU/host.

The NPU consists of 4 sub-neuron cores, routers, MUX and combinational circuits and adopts a crossbar array architecture to generate a local partial sum. To minimize routing congestion and power consumption due to the large clock tree of multicore SNN processor, we employed hierarchical layout with each NPU as a hard macro. As shown in Fig. 4(a), each NPU occupies an area of 1.02 mm^2 in 40nm CMOS technology node with the NoC routers placed at the center to ease the communication with multiple SRAM-based synaptic memories and the nearby sub-cores. Ultra-low leakage SRAM arrays are chosen to minimize the overall chip leakage. Area breakdown of the NPU circuitry is exhibited in Fig. 4(a), with 80% of the area occupied by the SRAM-based synapse memories.

IV. Measurement Results

A 12-NPU prototype chip has been fabricated in 40nm CMOS technology, occupies a total area of 19.76 mm^2, including I/O pads. Fig. 4(b) depicts the inter-connectivity of our neuromorphic processor with the FPGA-based host CPU via a μBlaze subsystem that incorporates a BRAM, control logic and a CPU–SNN bridge. The bridge is built to transfer data and instructions to activate the chip. Fig. 5 shows die micrograph, chip summary and measurement setup. Fig. 6 exhibits the dependency of accuracy and weight precision for the MNIST-MLP task. We adopt 5b weight value for a satisfactory accuracy ($> 96\%$) with minimal hardware cost. Besides, the architecture is demonstrated with a shape sensing application where an 8-core mapping can successfully recognize 5 different shapes (i.e., container, TV remote, laptop, toy gun, cup) with an accuracy of 97.8% by leveraging AI-on-Skin technique [9] (Fig. 7). Fig. 8 shows that the chip is fully functional down to 0.42V at 5KHz while the leakage and the dynamic power against VDD are shown in Fig. 9 and Fig. 10, respectively. The test chip achieves a minimum energy per



Technology	40nm ULP	
Network Type	SNN	
#Neurosynaptic Density	256 x 256 (5bit synaptic weight)	
Core Area	0.79 mm x 1.29 mm	
Die Area	4.00 mm x 4.94 mm	
V_{DD} (V)	0.5 V	0.9V
Leakage (µW)	100	500
Power (mW)	0.50	10.26
Frequency	12.5 MHz	100 MHz
Energy/SOP	1.71pJ	4.38 pJ
Energy/step (MNIST)	62.9 nJ	161.42 nJ

Fig. 5. Die micrograph, chip summary and measurement setup.

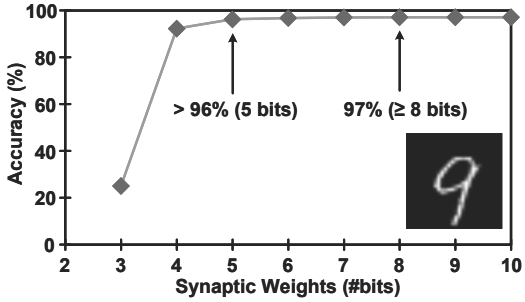


Fig. 6. Accuracy on MNIST dataset for synaptic weight bitwidth sweeping from 3 bits to 10 bits.

spiking operation of 1.7pJ at 0.5V. Typical output waveforms

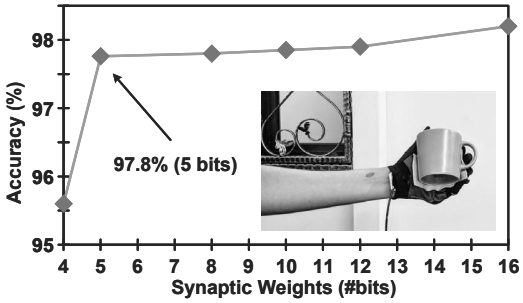


Fig. 7. Accuracy on AI-on-Skin (i.e.:shape sensing) for synaptic weight bitwidth sweeping from 4 bits to 16 bits.

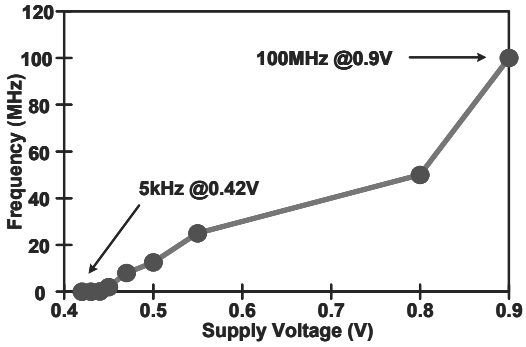


Fig. 8. Maximum frequency for different supply voltages ranging from 0.4V to 0.9V.

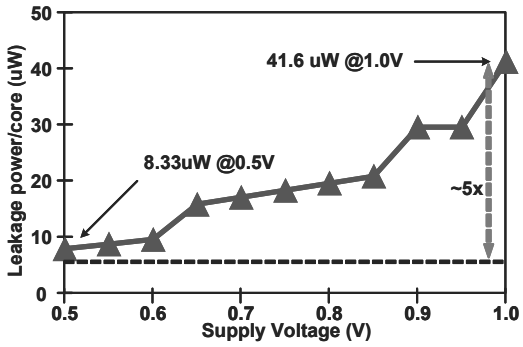


Fig. 9. Leakage power for different supply voltages ranging from 0.5V to 1.0V.

between the processor and the FPGA host at 0.5V are also captured and shown in Fig. 11.

Finally, we benchmark our test chip with the existing state of the art. As the table shows, our design consumes 1.7pJ/SOP, and 62.9nJ/step for the MNIST task, achieving 19% and 17% improvement compared to the reported best metrics, respectively. This is primarily due to our in-network computing circuits that eliminate extra core usage, the lightweight NoCs that remove the routing computation and the low power techniques that gate the inactive memory banks and routers.

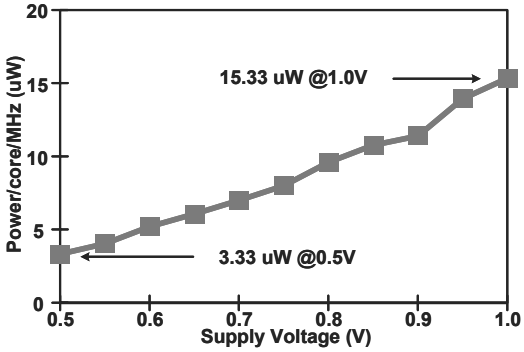


Fig. 10. Dynamic power for different supply voltages ranging from 0.5V to 1.0V.

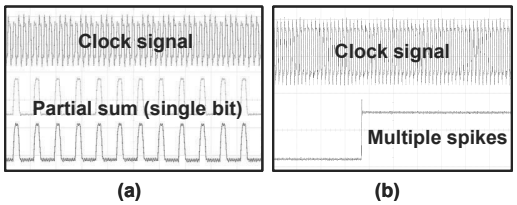


Fig. 11. (a) Partial sum single bit signal transmitted on partial sum router and received via north and south ports. (b) Multiple spikes transmitted on spike router.

Work	This Work	ASSCC'20 [1]	TCASI'21 [2]	ESSCIRC'21 [3]	ISSCC'22 [6]
Tech (nm)	40	40	40	40	28
Network Type	SNN	SNN	SNN	SNN	RNN
#Synapses (width)	64k x 12 (5-bit)	16k x 16 (8-bit)	64k x 16 (4-bit)	64k (4-bit)	132k (8-bit)
#Neurons	256 x 12	128 x 16	256 x 16	256 x 43	256-256-16
Chip Area (mm2)	19.76	4.80	5.28	19.66	0.86
Core Area (mm2)	1.11	0.30	0.33	0.25	0.45
Voltage (V)	0.5 - 0.9	0.5 - 1.0	0.85	0.5 - 1.0	0.5-0.8
Freq (MHz)	12.5-100	14.5 - 100	100	12.5 - 160	13-115
Condition	0.50V, 12.5MHz	0.5V, 14.5MHz	0.85V, 100MHz	0.5V, 12.5MHz	0.5V, 13MHz
Power (mW)	0.50	1.16	0.47	0.13	268.00
Energy/SOP (pJ)*	**1.71**	**4.80**	**4.50**	**2.10**	**7.57**
Energy/step (nJ)* (Application)	**62.90 (MNIST)**	**172.80 (MNIST)**	**162.00 (MNIST)**	**75.60 (MNIST)**	**50.00 (DVS Gesture)**
Dataset/Network (Accuracy %)	MNIST (96.11%)	MNIST (97.8%)	MNIST (97.9%)	MNIST (97%)	DVS Gesture (87.3%)
	Shapes Recog. (97.8%)			DVS Gesture (>80%)	KWS (90.7%)
	Single Character Recog. (96.7%)				Navigation (96.4%)

*** Energy metrics are normalized to 40nm for all designs.**
From [2], the work is synthesized using 40nm tech node.

Fig. 12. Benchmark table comparing this work with state-of-the-art.

V. Conclusion

In this work, we propose and fabricate a 12-NPU neuromorphic processor with In-Network Computing scheme to eliminate extra core usage and obtain lossless mapping on SNN. Assisted with low power circuit techniques, the chip achieves an accuracy of 96.1% on MNIST-MLP task, with an energy efficiency of 1.7pJ/SOP at 0.5V, 19% lower than the state-of-the-art result. This manifests that our processor is highly suitable for Artificial Intelligence applications on energy-constrained edge devices.

References

[1] V. P. Nambiar, J. Pu, Y. K. Lee, A. Mani, T. Luo, L. Yang, E. K. Koh, M. M. Wong, F. Li, W. L. Goh, and A. T. Do, "0.5V 4.8 pJ/SOP 0.93

μW Leakage/core neuromorphic processor with asynchronous NoC and reconfigurable LIF neuron," in *2020 IEEE Asian Solid-State Circuits Conference (A-SSCC)*, 2020, pp. 1–4.

[2] Junran Pu, Wang Ling Goh, Vishnu P. Nambiar, Ming Ming Wong, and Anh Tuan Do, "A 5.28-mm² 4.5pJ/SOP energy-efficient spiking neural network hardware with reconfigurable high processing speed neuron core and congestion-aware router," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 68, no. 12, pp. 5081–5094, 2021.

[3] M. M. Wong, S. B. Shrestha, V. P. Nambiar, A. Mani, Y. K. Lee, E. K. Koh, W. Jiang, K. T. C. Chai, and A. T. Do, "A 2.1 pJ/SOP 40nm snn accelerator featuring on-chip transfer learning using Delta STDP," in *ESSCIRC 2021 - IEEE 47th European Solid State Circuits Conference (ESSCIRC)*, 2021, pp. 95–98.

[4] Jeongwoo Park, Juyun Lee, and Dongsuk Jeon, "7.6 a 65nm 236.5 nJ/Classification neuromorphic processor with 7.5% energy overhead on-chip learning using direct spike-only feedback," in *2019 IEEE International Solid- State Circuits Conference - (ISSCC)*, 2019, pp. 140–142.

[5] Huiyu Mo, Wenping Zhu, Wenjing Hu, Guangbin Wang, Qiang Li, Ang Li, Shouyi Yin, Shaojun Wei, and Leibo Liu, "9.2 a 28nm 12.1 TOPS/W dual-mode cnn processor using effective-weight-based convolution and error-compensation-based prediction," in *2021 IEEE International Solid-State Circuits Conference (ISSCC)*, 2021, vol. 64, pp. 146–148.

[6] Charlotte Frenkel and Giacomo Indiveri, "Reckon: A 28nm sub-mm² task-agnostic spiking recurrent neural network processor enabling on-chip learning over second-long timescales," in *2022 IEEE International Solid-State Circuits Conference (ISSCC)*, 2022, vol. 65, pp. 1–3.

[7] Gregory K. Chen, Raghavan Kumar, H. Ekin Sumbul, Phil C. Knag, and Ram K. Krishnamurthy, "A 4096-neuron 1M-synapse 3.8PJ/SOP spiking neural network with on-chip STDP learning and sparse weights in 10nm FinFET CMOS," in *2018 IEEE Symposium on VLSI Circuits*, 2018, pp. 255–256.

[8] Bo Wang, Jun Zhou, Weng-Fai Wong, and Li-Shiuan Peh, "Shenjing: A low power reconfigurable neuromorphic accelerator with partial-sum and spike networks-on-chip," in *2020 Design, Automation Test in Europe Conference Exhibition (DATE)*, 2020, pp. 240–245.

[9] Ananta Narayanan Balaji and Li-Shiuan Peh, "AI-on-Skin: Enabling on-body AI inference for wearable artificial skin interfaces," in *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2021, CHI EA '21, Association for Computing Machinery.