1.7pJ/SOP, 0.5V Scalable Neuromorphic Processor with Integrated Partial Sum Router for In-Network Computing

B. Wang*[2], M. M. Wong*[1], D. Li[1,2], Y.S. Chong[1], J. Zhou[3], W. F. Wong[3], L. Peh[3], A. Mani[1], M. Upadhyay[3], A. Balaji[3], and A. T. Do[1]

*Equally contributed.

[1]Institute of Microelectronics, A*STAR, Singapore; [2]Singapore University of Technology and Design, Singapore; [3]Department of Computer Science, National University of Singapore.

Recent Spiking Neural Network (SNN) processors have successfully demonstrated impressive performance in digit recognition [1], object detection [2], robotic control [3] and event-based perception [4]. They leverage neuron cores for synaptic weight accumulation while deploying Networks-on-Chip (NoCs - routers & wire links) for spike delivery across the cores [5-7]. This leads to **(1)** information lost due to spike quantization and **(2)** area and power overheads due to the need for split-and-merge scheme to map a layer with large number of neurons on a SNN hardware (Fig. 1 (a), existing solutions [1,2]).

To overcome these issues, we have proposed in-network computing (INC) in Shenjing architecture [8] with compute-enabled NoCs to achieve lossless, flexible and energy-efficient SNN operations (Fig. 1 (b)). It **(1)** prevents the accuracy loss due to within-core quantization by enforcing partial sums to aggregate within the routers until a full total sum is formulated and **(2)** allows efficient core mapping with less physical cores (Fig. 1 (a)). Moreover, our software-defined NoC circuitry offloads the routing computation and the flow control from the processor. Combined with local hibernation techniques, the chip consumes only 0.5 mW power and 1.7pJ/SOP energy with lossless mapping from the algorithm to the hardware.

Fig. 2 (a) details the mapping of an illustrative 768×512×10 SNN onto Shenjing for MNIST digit inference. Layer 1 (i.e., L1 - 768×512) is mapped to the first 8 cores (i.e., red color 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 color). Note that we still require multiple cores to map a large layer, but the number of cores is 1/3 lesser than split-and-merge scheme [1,2], thanks to the partial sum router. Considering L1, instead of quantizing the spikes to Yes/No within each core, the partial sum router integrates an adder for accumulation along the data propagation until it arrives at core (0,0) and core (0,1) where the final spikes are sent to L2. Eventually, the data flow forms an adder tree, starting from the 8 cores and aggregating at the root node core (0,2) where the classification is made (Fig. 2 (b)). 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 spike within each core. Fig. 2 (c) depicts the circuit diagram of a Neural Process Unit (NPU), including a neuron core, a partial sum router and a spike router. The neuron core is implemented with SRAM banks with the multiplication for synaptic weight translated into memory read. The local partial sum from each neuron can be injected to the partial sum NoC 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.

Low power circuit techniques are incorporated in the Shenjing SNN chip to further reduce its power and to 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 2 vertical cores can be operated in parallel to improve 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 is 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 bit is set and programmed to the chip, 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 promotes efficient core mapping and effective energy breakdown by decomposing a complex task into a population of atomic operations. An example is illustrated in Fig. 3 (d).

Our routers implementation is illustrated in Fig. 4(a). 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 or bypasses the incoming spike as a crossbar switch between the input and the output ports. All the routing operations are defined by the instructions that are preloaded in a tiny register file and a LUT, depending on the network structure and the connectivity between the cores. The neuron core consists of sub-cores, MUX and combinational circuits and adopts a crossbar array architecture [2] 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 IP. 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 cells 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.

A 12-NPU Shenjing prototype 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 interconnectivity 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, instructions as well as to activate the Neuromorphic Computing (NC) chip. Fig. 7 shows the die micrograph, measurement setup and our chip's specification.

Fig. 5 (a) exhibits the dependency of accuracy and weight precision for the MNIST-MLP task. We adopt 5b weight values for a satisfactory accuracy (> 96%) with minimal hardware cost. Besides, the Shenjing 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. 5 (b)). Fig. 5 (c) shows that the chip is fully functional down to 0.42V at 5 KHz while the leakage and the dynamic power against VDD are shown in Fig. 5 (d) and (e), respectively. The test chip achieves a minimum energy per spiking operation of 1.7pJ at 0.5V. Typical output waveforms between Shenjing and FPGA host at 0.5V were also captured and shown in Fig. 6.

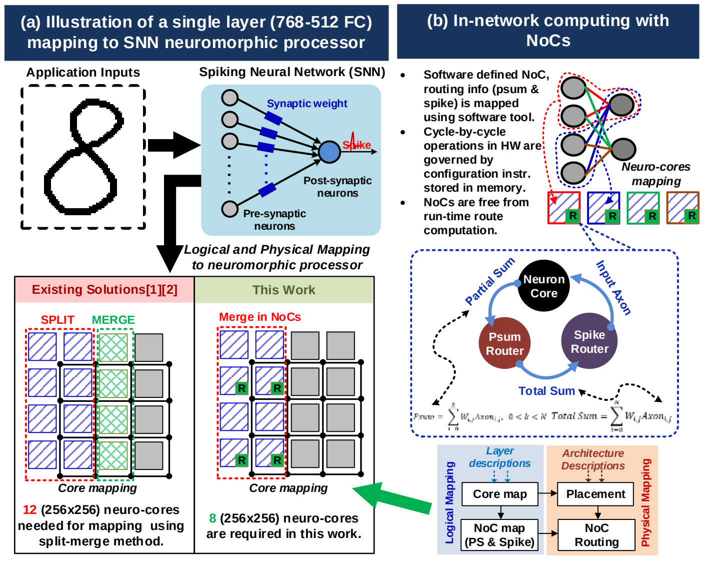
Finally, we benchmark our test chip with the existing state-of-the-art. As the table shows, our design consumes 1.7pJ/SOP, and 62.9pJ/step for the MNIST task, achieving 18.6% and 16.8% improvement compared to the reported best metrics, respectively. This is primarily due to our in-network computing circuits 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.

Acknowledgements:

The work was partially supported by NRF, Singapore (NRF-CRP23-2019-0003) and SUTD, Singapore (SRG-ISTD-2020-162, SGP-AIRS1841).

References:

[1] V. P. Nambiar et al., A-SSCC 2020, pp. 1-4.
[2] F. Akopyan et al., in IEEE TCAD, 2015, pp. 1537-1557.
[3] J. Dupeyroux et al., ICRA 2021, pp. 96-102.
[4] C. Frenkel et al., ISSCC 2022, pp. 1-3.
[5] J. Pu et al., in IEEE TCAS–I, 2021, pp. 5081-5094.
[6] M. M. Wong et al., ESSCIRC 2021, pp. 95-98.
[7] G. K. Chen et al., in JSSC, 2019, pp. 992-1002.
[8] B. Wang et al., DATE 2020, pp. 240-245.
[9] A. Balaji et al., CHI Extended Abstracts 2021, pp. 1-7.


(a) Illustration of a single layer (768-512 FC) mapping to SNN neuromorphic processor
Application Inputs
Spiking Neural Network (SNN)
Synaptic weight
Spike
Post-synaptic neurons
Pre-synaptic neurons
Logical and Physical Mapping to neuromorphic processor
Existing Solutions[1][2]
This Work
SPLIT
MERGE
Merge in NoCs
R
Core mapping
Core mapping
12 (256x256) neuro-cores needed for mapping using split-merge method.
8 (256x256) neuro-cores are required in this work.
(b) In-network computing with NoCs
Neuro-cores mapping
Neuron Core
Partial Sum
Input Axon
Psum Router
Spike Router
Total Sum
Layer descriptions
Architecture Descriptions
Core map
Placement
NoC map (PS & Spike)
NoC Routing
Logical Mapping
Physical Mapping

- Software defined NoC, routing info (psum & spike) is mapped using software tool.
- Cycle-by-cycle operations in HW are governed by configuration instr. stored in memory.
- NoCs are free from run-time route computation.

$$P_{sum} = \sum_{i=0}^{k} W_{i,j} Axon_{i,j},\ 0 < k < N \quad Total\ Sum = \sum_{i=0}^{N} W_{i,j} Axon_{i,j}$$

Fig. 1. (a) Core-saving of Shenjing thanks to the proposed INC concept which includes (b) partial-sum & spike routers.

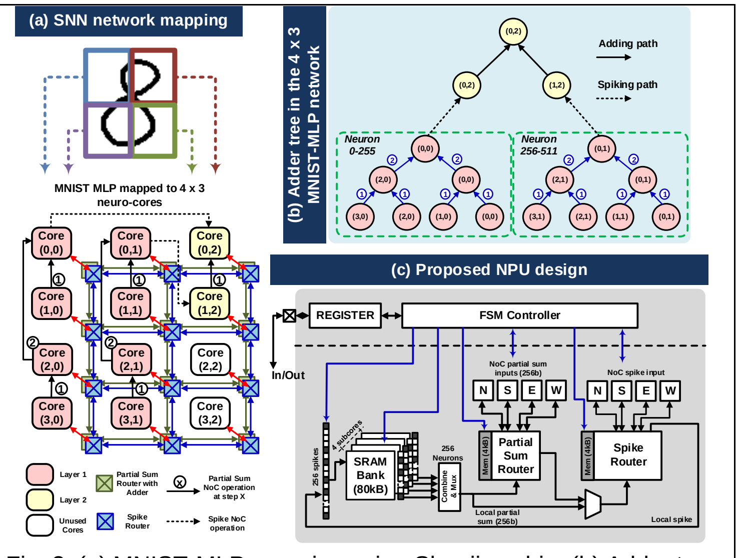
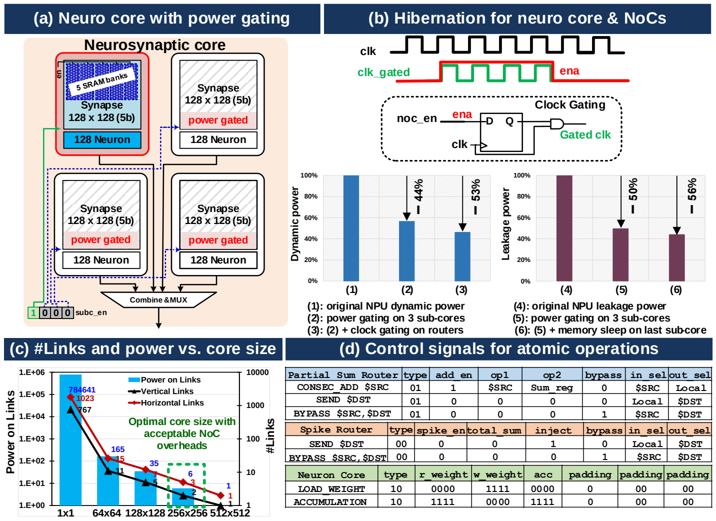


Fig. 2. (a) MNIST-MLP mapping using Shenjing chip. (b) Adder tree structure is pre-mapped to partial sum addition to eliminate accuracy loss. (c) Detailed architecture of the proposed NPU design.

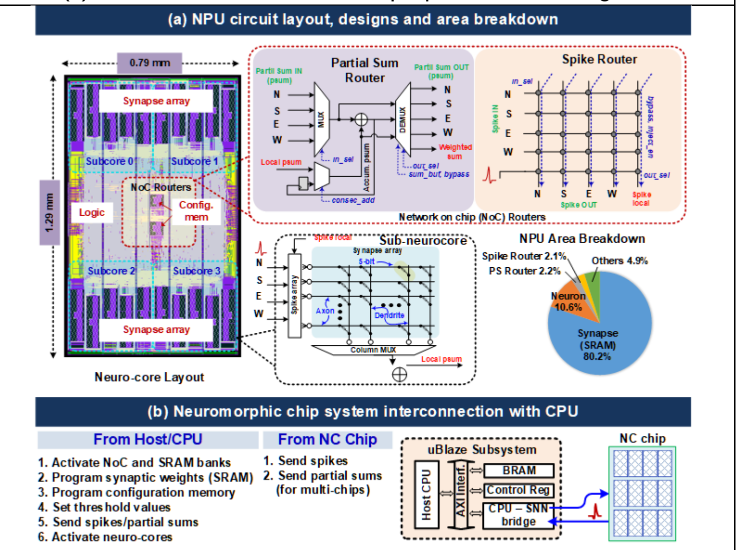

(a) Neuro core with power gating
Neurosynaptic core
5 SRAM banks
Synapse 128 x 128 (5b)
128 Neuron
power gated
Combine &MUX
subc_en
1 0 0 0
(b) Hibernation for neuro core & NoCs
clk
clk_gated
ena
noc_en
Clock Gating
D Q
Gated clk
Dynamic power
Leakage power
44%
53%
50%
56%
(1)
(2)
(3)
(4)
(5)
(6)
(1): original NPU dynamic power
(2): power gating on 3 sub-cores
(3): (2) + clock gating on routers
(4): original NPU leakage power
(5): power gating on 3 sub-cores
(6): (5) + memory sleep on last subcore
(c) #Links and power vs. core size
Power on Links
Vertical Links
Horizontal Links
Optimal core size with acceptable NoC overheads
Power on Links
#Links
1x1
64x64
128x128
256x256
512x512
(d) Control signals for atomic operations

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 activated individually according to the neurons needed. (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 neuro core and NoCs.



From Host/CPU

1. Activate NoC and SRAM banks
2. Program synaptic weights (SRAM)
3. Program configuration memory
4. Set thresholds values
5. Send spikes/partial sums
6. Activate neuro-cores

From NC Chip

1. Send spikes
2. Send partial sums (for multi-chips)

Fig. 4. (a) NPU layout and circuit designs for the sub-neuron core and NoCs with the area breakdown are shown. (b) Microblaze (FPGA) is used to connect Shenjing chip and CPU/host.

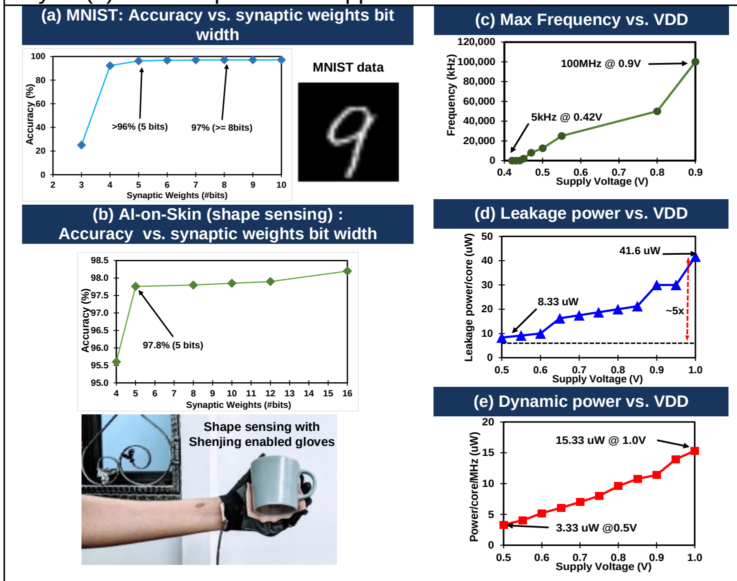
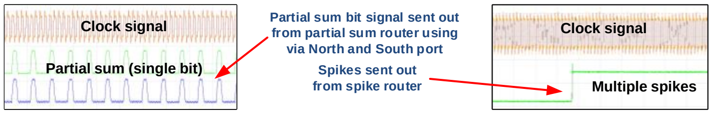


Fig. 5. Simulation inference accuracy for (a) MNIST and (b) AI-on-Skin applications and power and frequency measurements (c-e).

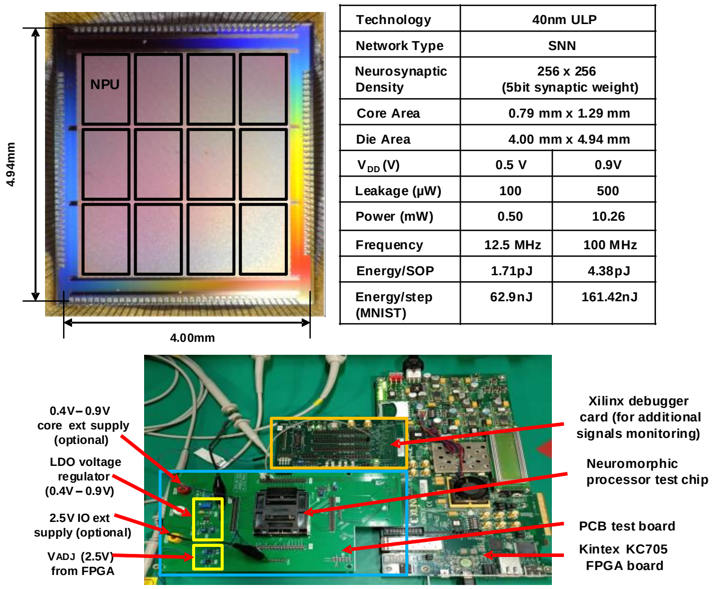

Clock signal
Partial sum (single bit)
Partial sum bit signal sent out from partial sum router using via North and South port
Spikes sent out from spike router
Clock signal
Multiple spikes

Measurement observation from partial sum and spiking chip testing

Work	This Work	ASSCC'20 [1]	ISSCC'22 [4]	TCASI'21 [5]	ESSCIRC'21 [6]	JSSC'19 [7]
Tech (nm)	40	40	28	40	40	10
Network Type	SNN	SNN	RNN	SNN	SNN	SNN
#Synapses (width)	64k x 12 (5-bit)	16k x 16 (8-bit)	132k (8-bit)	64k x 16 (4-bit)	64k (4-bit)	1M (7-bit)
#Neurons	256 x 12	128 x 16	256-256-16	256 x 16	256 x 43	64 x 64
Chip Area (mm2)	19.76	4.80	0.86	5.28	19.66	1.72
Core Area (mm2)	1.11	0.30	0.45	0.33	0.25	0.027
Voltage (V)	0.5 - 0.9	0.5 - 1.0	0.5-0.8	0.85	0.5 - 1.0	0.5 - 0.9
Freq (MHz)	12.5-100	14.5 - 100	13-115	100	12.5 - 160	105 - 506
Condition	0.50V, 12.5MHz	0.5V, 14.5MHz	0.5V, 13MHz	0.85V, 100MHz	0.5V, 12.5MHz	0.5V, 105MHz
Power (mW)	0.50	1.16	268.00	0.47	0.13	6.20
Energy/SOP (pJ)*	**1.71**	**4.80**	**7.57**	**4.50**	**2.10**	**15.20**
Energy/step (nJ)* (Application)	**62.90 (MNIST)**	**172.80 (MNIST)**	**50.00 (DVS Gesture)**	**162.00 (MNIST)**	**75.60 (MNIST)**	**4000.00 (MNIST)**
Dataset/Network (Accuracy %)	MNIST (96.11%) Shape Recog. (97.8%) Single Character Recog. (96.7%)	MNIST (97.8%)	DVS Gesture (87.3%) KWS (90.7%) Navigation (96.4%)	MNIST (97.9%)	MNIST (97%) DVS Gest (>80%)	MNIST (98%)

* Energy metrics are normalized to 40nm for all designs.
From [5], the work is synthesized using 40nm tech node.

Fig. 6. Measurement observations at 0.5 V supply voltage and comparison with state-of-the-art.


NPU
4.94mm
4.00mm
0.4V – 0.9V core ext supply (optional)
LDO voltage regulator (0.4V – 0.9V)
2.5V IO ext supply (optional)
VADJ (2.5V) from FPGA
Xilinx debugger card (for additional signals monitoring)
Neuromorphic processor test chip
PCB test board
Kintex KC705 FPGA board

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	4.38pJ
Energy/step (MNIST)	62.9nJ	161.42nJ

Fig. 7. Die micrograph, chip summary and measurement setup.