IDEC Chip Design Contest

A 28nm Digital Neuromorphic Processor with Reconfigurable Neurons and Memory-Efficient Event Routing

ChangMin Ye*, Choongseok Song*, and Doo Seok Jeong*
*Division of Materials Science and Engineering, Hanyang University, Seoul, Republic of Korea
{skar928, dooseokj} @hanyang.ac.kr

Introduction

- Spiking neural networks (SNNs) process time-dependent data by capturing spatio-temporal patterns through the dynamic behavior of spikes in neurons and synapses.
- State variables in spiking neurons are updated in response to incoming presynaptic spikes, which are subsequently translated into output spike events.
- The primary computation in SNNs, known as a synaptic operation (SynOP), is triggered only when a spike occurs and utilizes locally stored information.

➡ **Due to these event-driven computations and the inherent sparsity of spike activities, SNNs naturally achieve low power consumption and are regarded as energy-efficient models.**

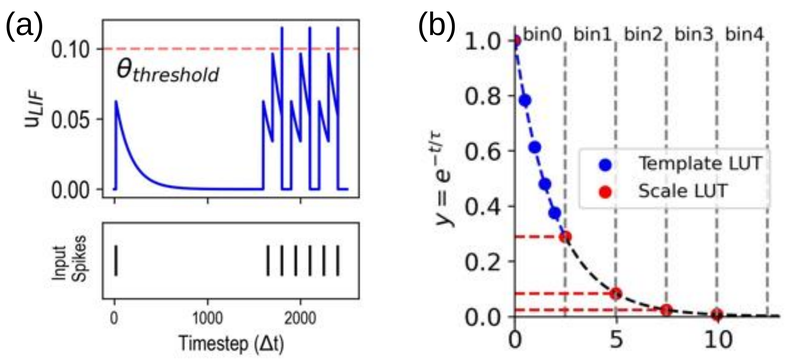
- Neuromorphic processors capitalize on these properties by executing SynOPs in an event-based manner across many processing cores.
- Their architectures typically feature distributed on-chip memories, which act as localized data stores to support spike-triggered SynOPs.

➡ **Consequently, designing efficient memory usage strategies becomes essential in neuromorphic systems, given the distributed nature of memory and the limited storage available per core.**

Memory efficient digital based Neuromorphic Processor

Efficient state update via TS-EFA

- Most spiking neuron models incorporate temporal kernels based on exponential functions, with the leaky integrate-and-fire (LIF) model being a typical example, as illustrated in Fig. 1(a).
- The template-scaling-based exponential function approximation (TS-EFA) technique [1] enables compact and efficient computation of exponentials using lightweight lookup tables (LUTs), as depicted in Fig. 1(b).
- In the proposed processor, TS-EFA is adopted to update the state variables of spiking neurons with minimal memory overhead.



Template LUT		Scale LUT	
Address	Data	Address	Data
0	1	0	1
1	0.781	1	0.297
2	0.623	2	0.081
⋮	⋮	⋮	⋮

Fig. 1. (a) Description of the LIF neuron model. (b) Template and scale LUTs for TS-EFA.

Group-wise Event Routing Mechanism

- Conventional event-routing in neuromorphic processors typically relies on large neuron-centric LUTs, which map presynaptic neurons to their corresponding postsynaptic targets and associated synaptic weights, as illustrated in Fig.2(a).
- In contrast, the group-wise event-routing (GWER) architecture [2] eliminates the need for such large LUTs by computing the addresses of postsynaptic neurons dynamically, as shown in Fig. 2(b).
- The necessary routing parameters for GWER are compactly stored in lightweight, layer-centric LUTs, enabling more memory-efficient communication.

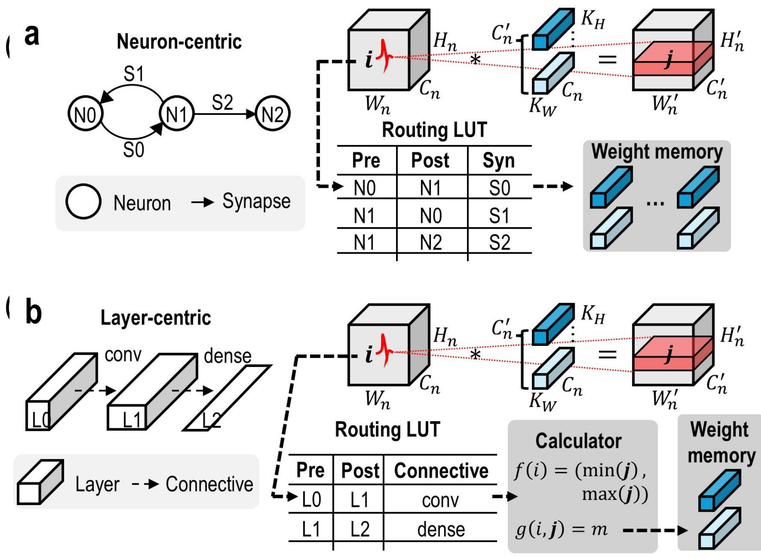


Fig. 2. The routing LUT for (a) neuron-centric and (b) layer-centric methods.

System Architecture and Event Processing flow

- Step1) Incoming off-chip events are delivered in packet form using the Address Event Representation (AER) protocol.
- Step2) Theses packets are routed to the neuromorphic core, where they are temporarily stored in an input buffer.
- Step3) The postsynaptic addresses corresponding to the buffered events are dynamically computed using the GWER method.
- Step4) In the neuron block, the state variables are updated using the TS-EFA technique to efficiently realize exponential dynamics.

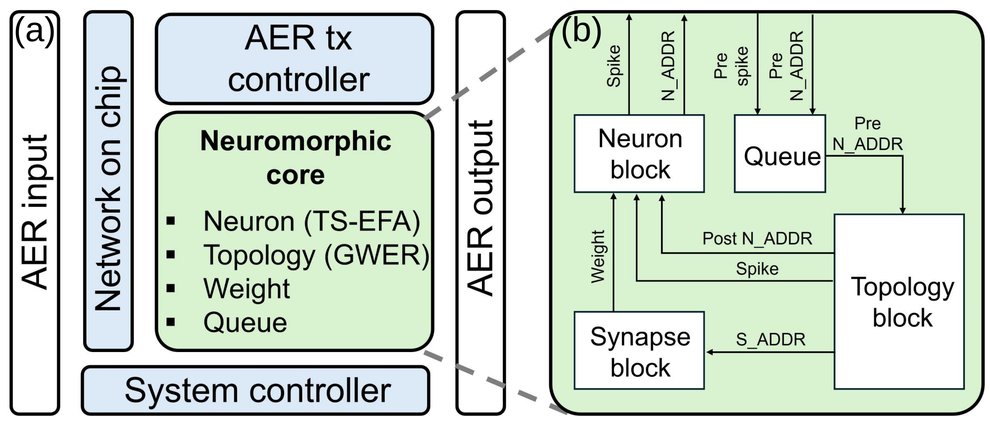
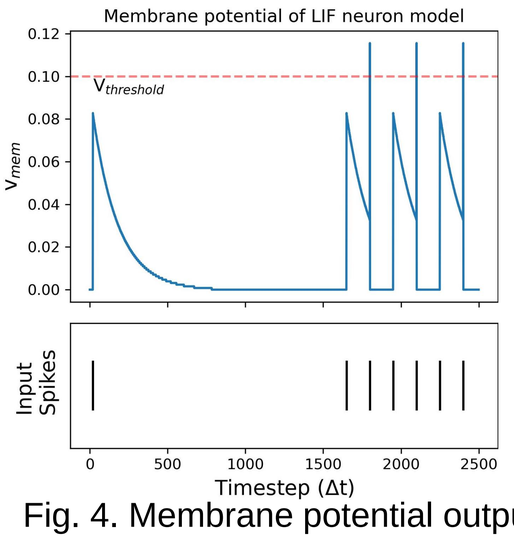


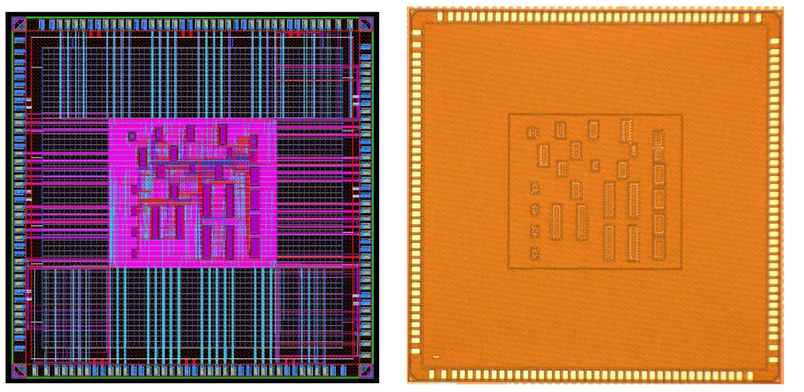
Fig. 3. Block diagram of (a) the overall architecture of the proposed processor and (b) the detailed architecture of the neuromorphic core.

Results



Parameter	Description	Value
Δt	Timestep width	2^{-12}s
τ_m	Membrane potential time-constant	0.04s
w	Synaptic weight	0.083
v_{thr}	Spiking threshold	0.1

Fig. 4. Membrane potential output result of the LIF neuron model configured on the chip (left) and its hyperparameters (right).



Chip summary	
Technology	28nm CMOS
Area	3.05 mm^2
Supply Voltage	1.0 V
Frequency	100 MHz
# neurons/core	2k
# synapses/core	32k
Routing mem.	18.4b/neuron
Neuron mem.	56b/neuron

Fig. 5. Chip layout (left), die photo (middle), and summary (right)

Conclusion

- We implemented a single-core neuromorphic processor in 28nm CMOS, integrating TS-EFA based spiking neuron dynamics [1] and LaCERA-based event-routing [2].
- The processor supports flexible neuron configurations, requiring only 56 bits of memory per neuron for reconfigurable spiking behavior.
- It also provides a scalable SNN mapping capability, with a compact memory footprint of 18.4 bits per neuron.

References

[1] Kim, Jeeson, et al. "Hardware-efficient emulation of leaky integrate-and-fire model using template-scaling-based exponential function approximation.“ IEEE Transactions on Circuits and Systems I: Regular Papers 68.1 (2020): 350-362.
[2] Ye, ChangMin, et al. "LaCERA: Layer-centric event-routing architecture.“ Neurocomputing 520 (2023): 46-59.

Acknowledgement

- The chip fabrication and EDA tool were supported by the IC Design Education Center (IDEC), Korea.