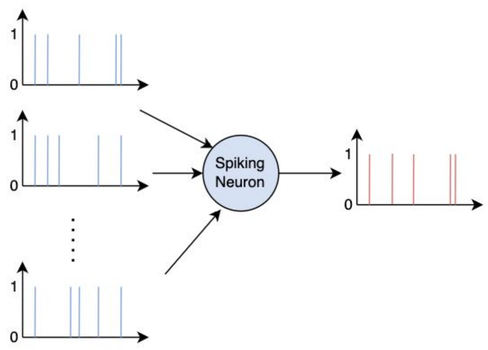
Evaluating Neuron Models for a 130-nm Spiking Neural Network Hardware Accelerator

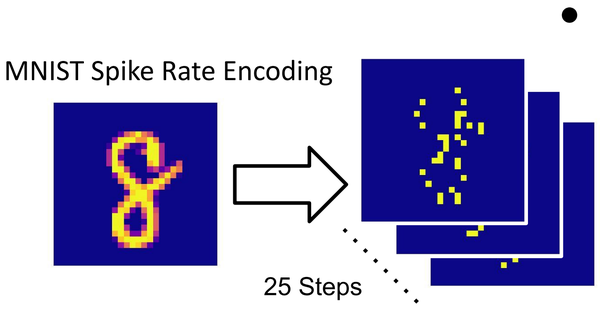
Amy Liu, Finn Moore, Kristen Schang, Isaac Schuster, Emily Su
University of Michigan

Project Motivation and Goal

Motivation

- **Spiking neural networks (SNNs)** encode information with temporal binary spikes, enabling power-efficient neural network models





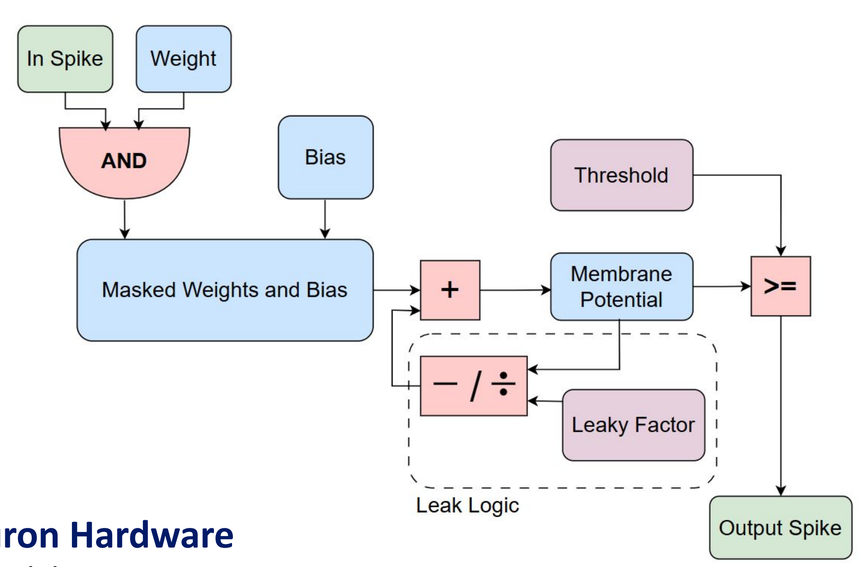
Goal

- Evaluate the efficiency of different neuron models
- Design simple, low-power SNN hardware to classify handwritten digits of the MNIST dataset
- SNNs are useful for edge AI applications that prioritize battery life (e.g., IOT, smart devices, embedded vision) without sacrificing performance

Neuron Models	Decay Type	Est. Accuracy	Est. Power
Integrate-and-Fire (IF)	N/A	High	Lowest
Leaky Integrate-and-Fire (LIF)	Exponential	Highest	Highest
Linear Decay LIF (LLIF)	Linear	High	Low

Spiking Neural Network Design and Implementation

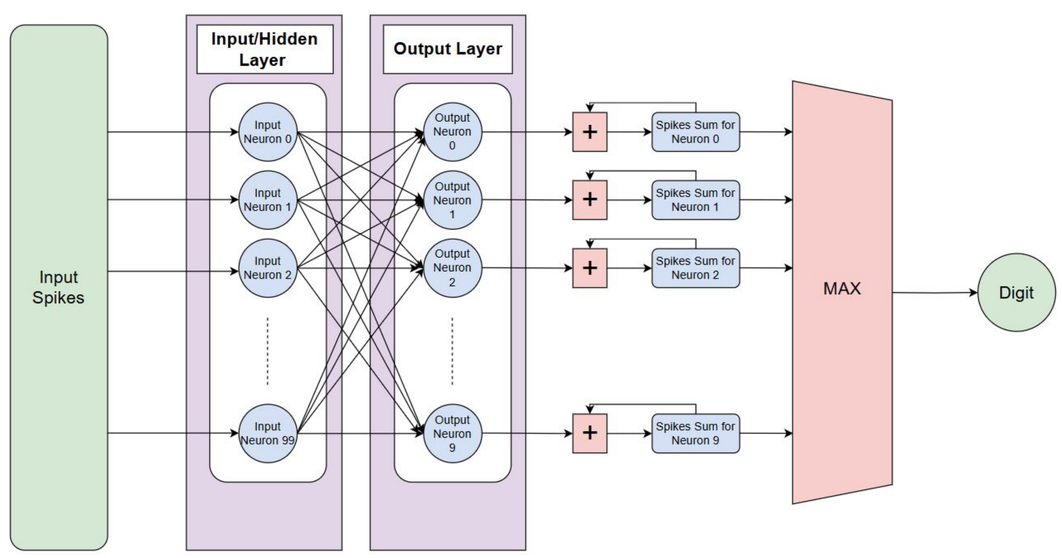
Neuron Architecture



Neuron Hardware

- Leak logic
  - IF: No leaking
  - LIF: divide stored membrane potential every cycle → exponential decay
  - LLIF: subtract from stored membrane potential every cycle → linear decay
- Weights and biases quantized to 8-bit fixed point

Full SNN Architecture



Architecture Specifications

- 28x28 flattened to 784 input spikes
- 2 fully connected layers: 100 and 10 neurons
- Input spikes are sent to the network one at a time, to each neuron in parallel

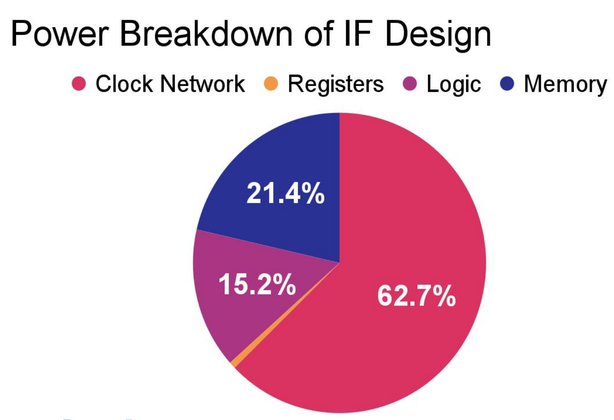
Results and Conclusion

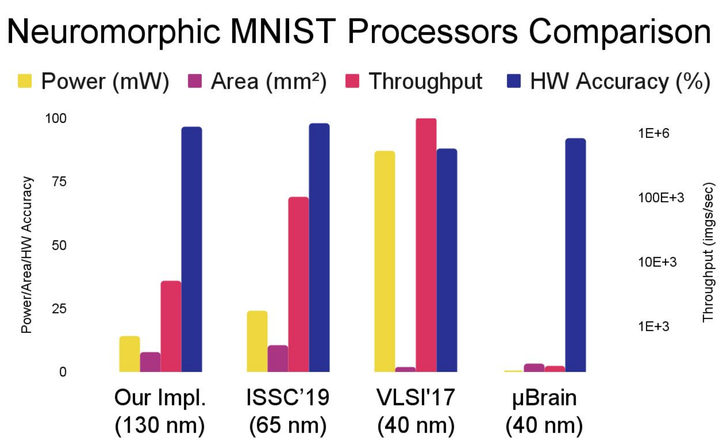
Testing Methodology

- Model weights trained using the snnTorch Python library on MNIST rate coded over 25 steps
- Simulated hardware models (IF, LIF, LLIF) with Synopsys VCS on 10,000 input streams
- Synthesized in IBM 130 nm process using Synopsys DC with ARM SRAM IP to store weights

Comparison with Other SNN Accelerators for MNIST

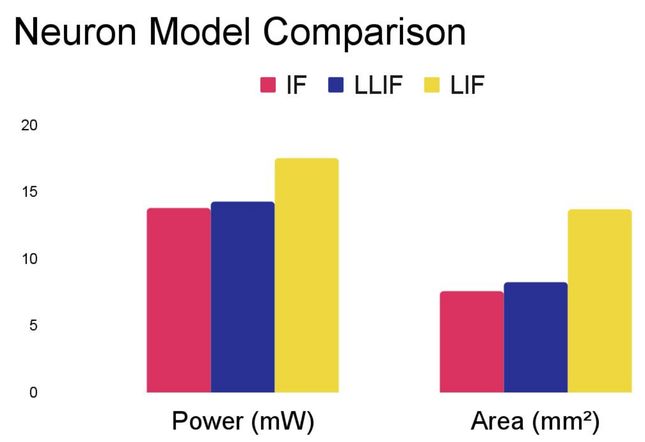
Implementation	Power	Area	Throughput	HW Accuracy	Tech
Ours (IF)	13.71 mW	7.51 mm^2	5K imgs/sec	96.28%	130 nm
ISSC'19 [1]	23.6 mW	10.08 mm^2	100K imgs/sec	97.83%	65 nm
VLSI'17 [2]	87.0 mW	1.31 mm^2	1.7M imgs/sec	88%	40 nm
µBrain [3]	73 µW	2.68 mm^2	238 imgs/sec	91.7%	40 nm





Neuron Model Comparison

Model	Power	Area	SW Accuracy
IF	13.71 mW	7.51 mm^2	97.37%
LIF	17.47 mW	13.66 mm^2	97.83%
LLIF	14.21 mW	8.16 mm^2	N/A



Conclusion

- Our design strikes a healthy balance between performance, efficiency, and accuracy
- Exponential leak logic creates significant overhead with marginal accuracy benefits; linear leak logic is more feasible

References

[1] J. Park, J. Lee et al, "A 65-nm Neuromorphic Image Classification Processor With Energy-Efficient Training Through Direct Spike-Only Feedback," in *IEEE Journal of Solid-State Circuits*, vol. 55, no. 1, pp. 108-119, Jan. 2020.
[2] F. N. Buhler et al, "A 3.43tops/w 48.9pj/pixel 50.1nj/classification 512 analog neuron sparse coding neural network with on-chip learning and classification in 40nm cmos," in *2017 Symposium on VLSI Circuits*, 2017, pp. C30–C31.
[3] J. Stuijt et al. "µbrain: An event-driven and fully synthesizable architecture for spiking neural networks," *Frontiers in neuroscience*, vol. 15, p. 664208, 2021.