MULTICORE/MANY-CORE SYSTEMS-ON-CHIP

Energy-Efficient Spiking Neural Networks Using Approximate Neuron Circuits and 3D Stacking Memory

School of Computer Science and Engineering
University of Aizu, Fukushima, Japan

Authors: Ryoji Kobayashi, Ngo-Doanh Nguyen, Nguyen Anh Vu Doan, Khanh N. Dang
E-mail: s1290176@u-aizu.ac.jp

Dec. 17, 2024

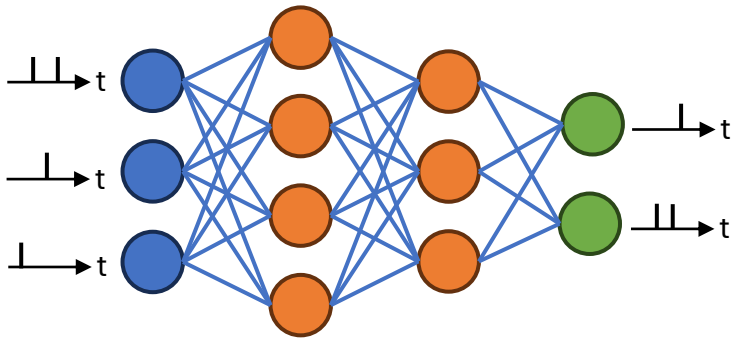
Outline

- Research Introduction
- Methodology
- Evaluation
- Conclusion

Outline

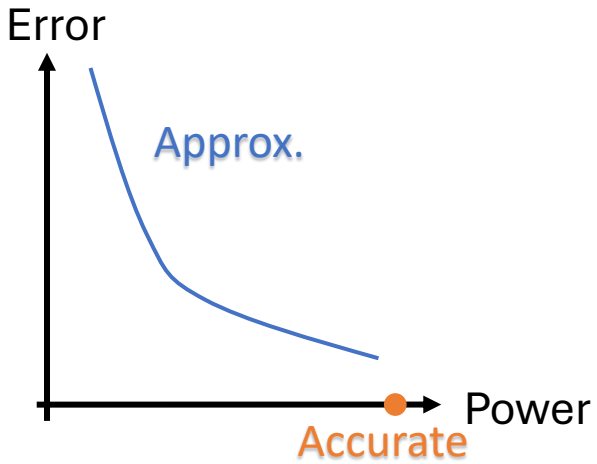
- Research Introduction
- Methodology
- Evaluation
- Conclusion

Research Introduction



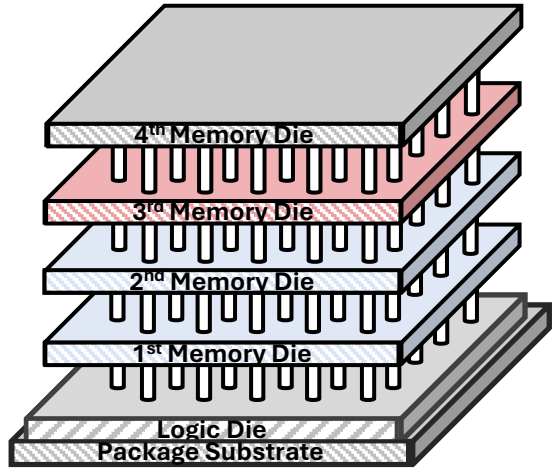
Spiking Neural Network

- Low power operation
- Simple neuron impl.
- Noise resilience



Approx. Computing

- Low power & latency computation
- HW design choice



3D Stacking Memory

- Scalability
- High-bandwidth
- Small footprint

➡ *Spiking Neural Net. with Approx. Neuron and 3D Stack. Memory*

Outline

- Research Introduction
- Methodology
- Evaluation
- Conclusion

Methodology (1/4)

- Approximate Computing
  - Allowance of accuracy loss -> Energy-efficient HW implementation
  - Effective for error-tolerant applications
    - Multimedia Processing
    - Machine Learning

However?

➡ ☹ *Produce noises in the computing* ➡ ☺ *Noise resilience of Spiking Neural Net.*

Our Contribution

- Approximate Spiking Neural Net. Implementation using:
  - Approx. Neurons
  - Approx. Memory

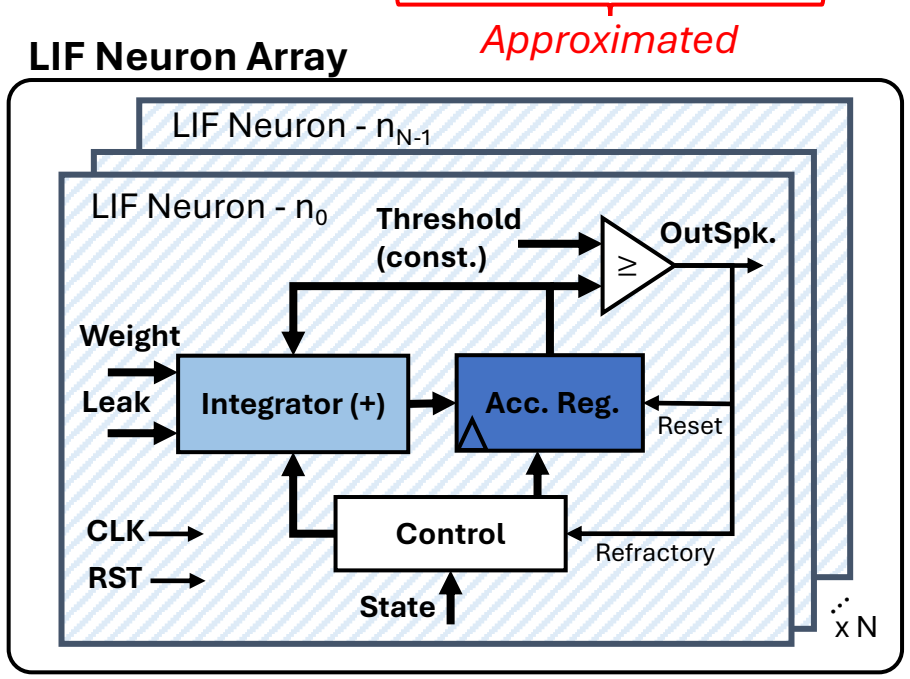
Methodology (2/4)

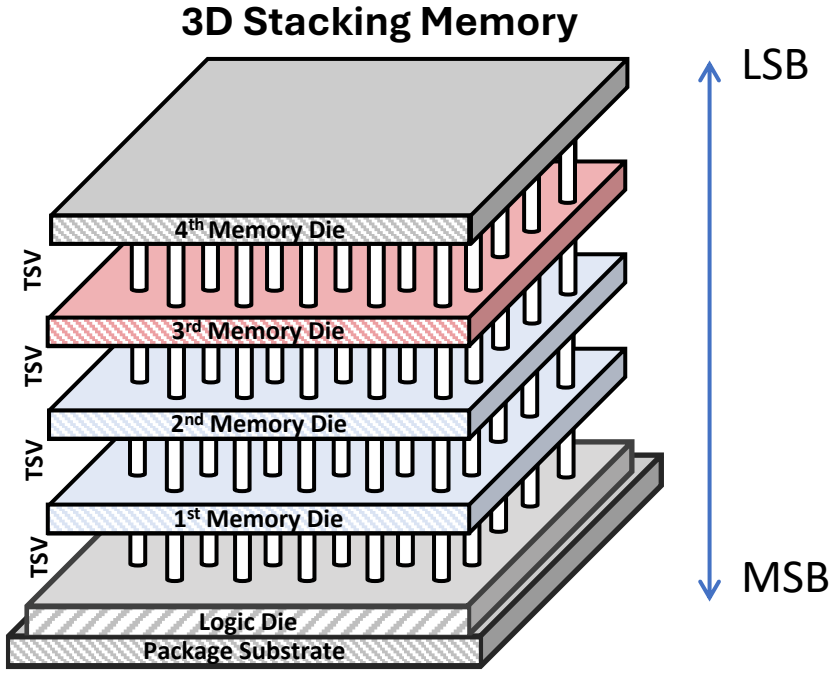
- Approximate Neuron
  - Use of Approx. Adder

$$V_i(t) = V_i(t-1) + \underbrace{\sum_j w_{ij}x_j(t-1) - \lambda}_{\textit{Approximated}}$$

- Approximate Memory
  - Undervolting
  - Power-gating

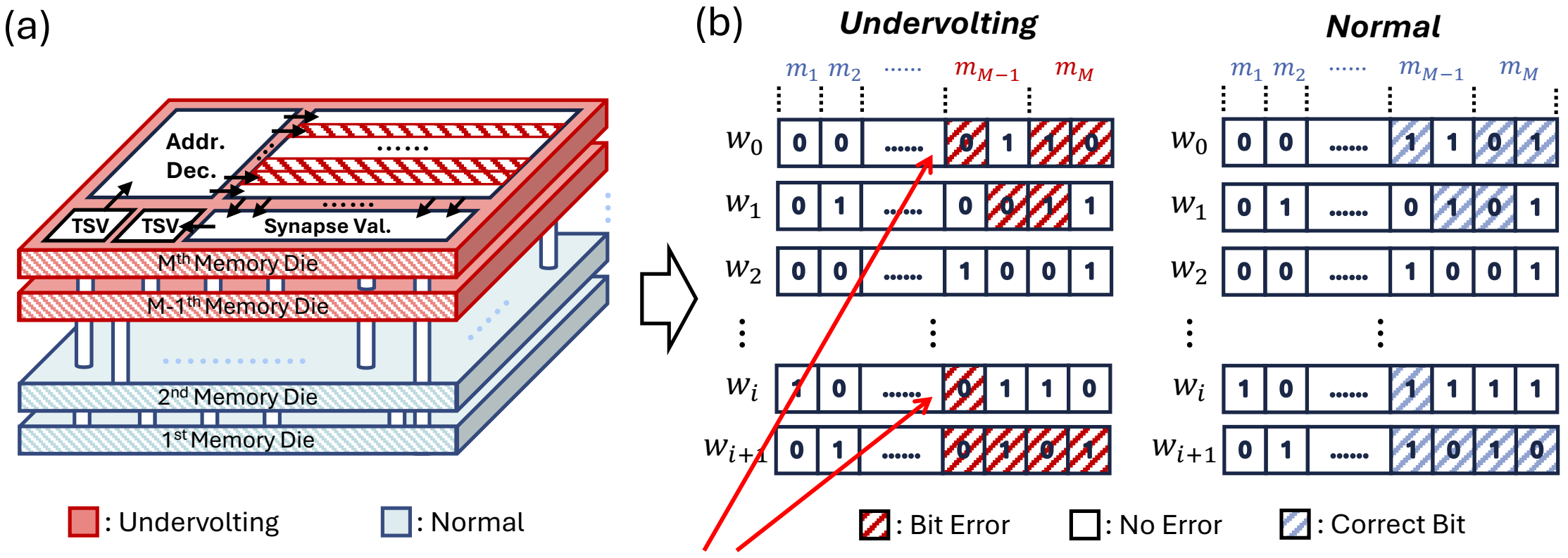
☺ *low power*
☹ *bit-error*





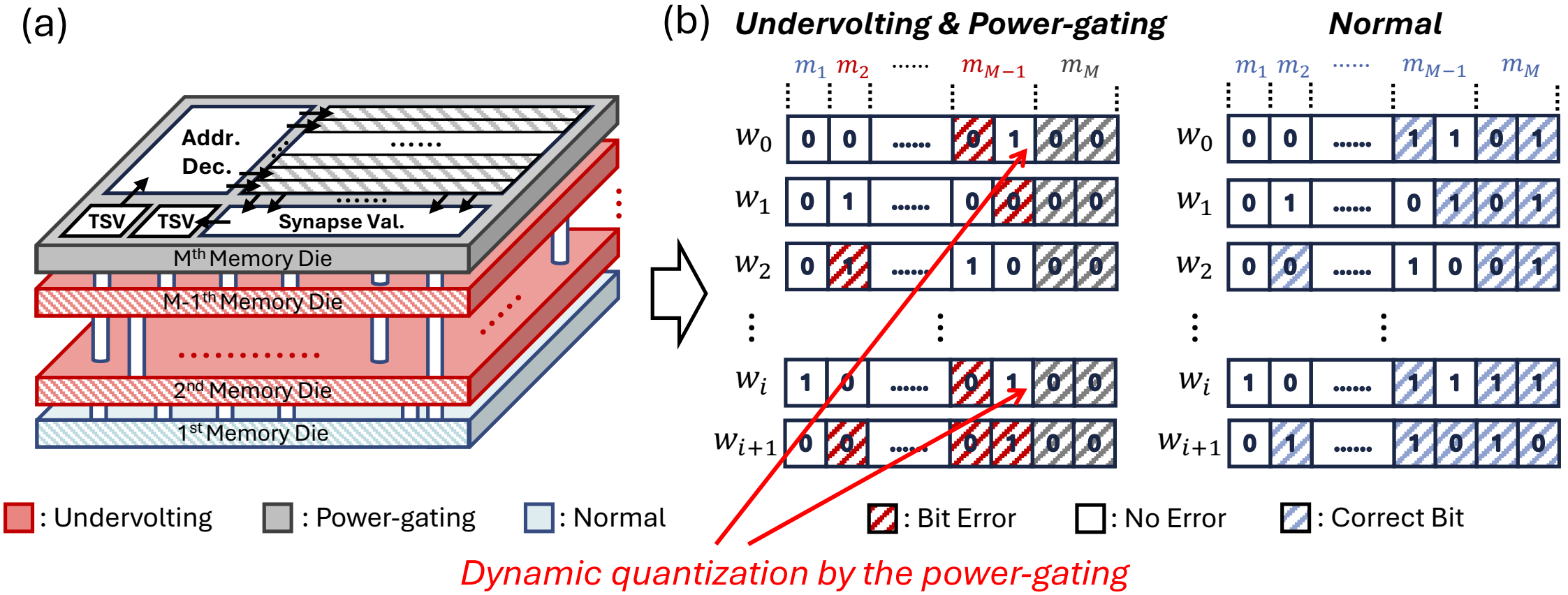
Methodology (3/4)

- Example of **Undervolting** operation in memory



Methodology (4/4)

- Example of **Undervolting & Power-gating** operation in memory

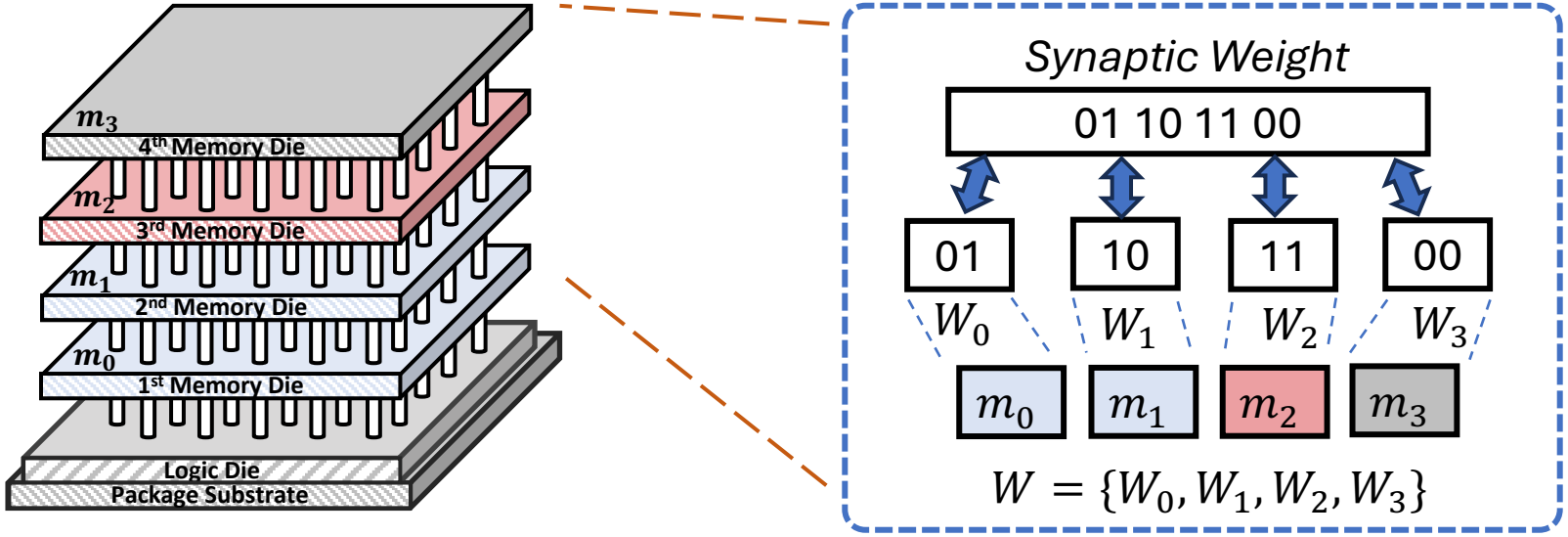


Outline

- Research Introduction
- Methodology
- Evaluation
- Conclusion

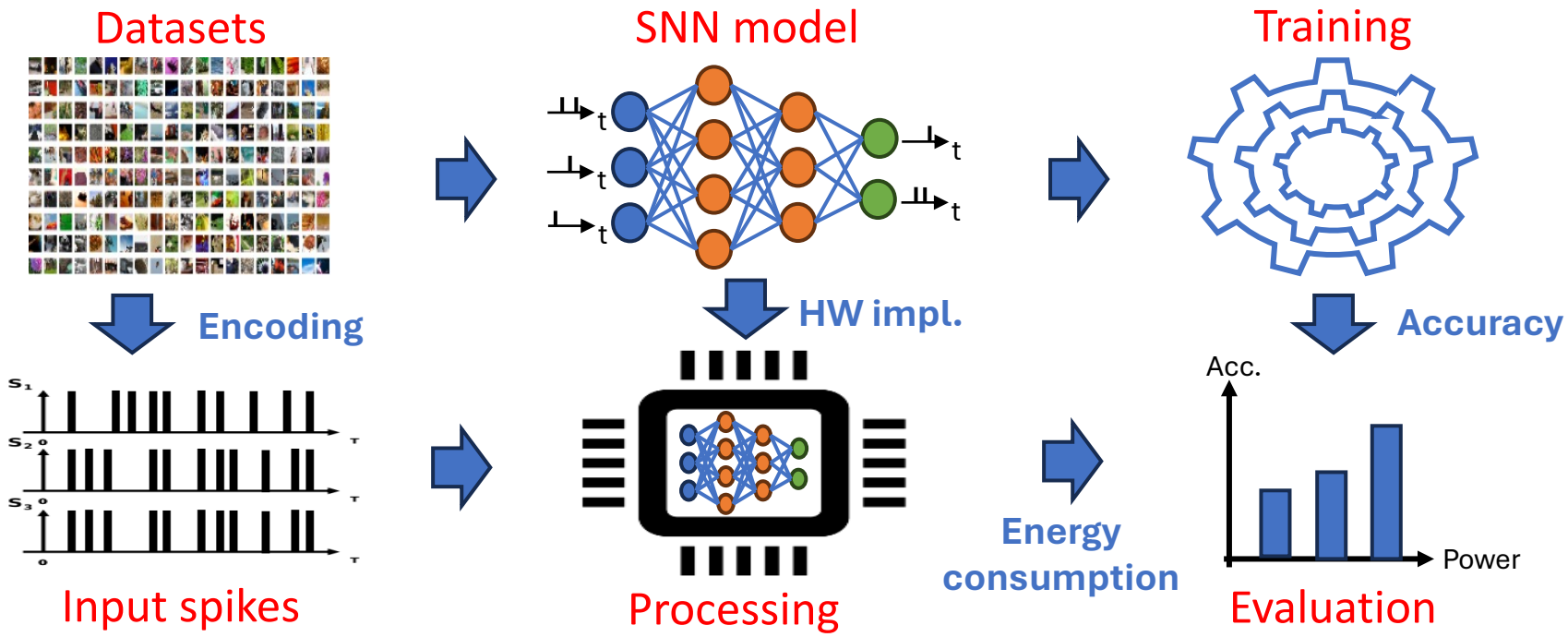
Evaluation: Setup (1/2)

- Memory Configurations
  - Split memory into 4 layers
  - Undervolting or power-gating each layer separately
- Synaptic weights are quantized to 8-bits and divided into 2-bits each



Evaluation: Setup (2/2)

- Spiking Neural Net. config. = [784 : 48 : 10]
  - Dataset: MNIST
- Power assessment tool = PrimeTime from Synopsys



Evaluation: Result (1/4)

- Normal Operation (Only use of Acc. or Approx. Adders [1])

Adder Name	Accuracy (%)	Energy per Neuron (nJ)	Area Reduction per Neuron (%)
ACC (Accurate)	94.8	4.039	-
5QT	94.3	3.976	1.59
5QC	94.0	3.942	2.93
5L8	94.1	3.036	9.94
5RP	93.3	3.192	17.31
5RL	93.5	3.328	17.13
5KB	94.5	3.604	9.41
5SV	83.0	3.668	10.67
5YE	41.4	2.919	24.14

Approx. Adders

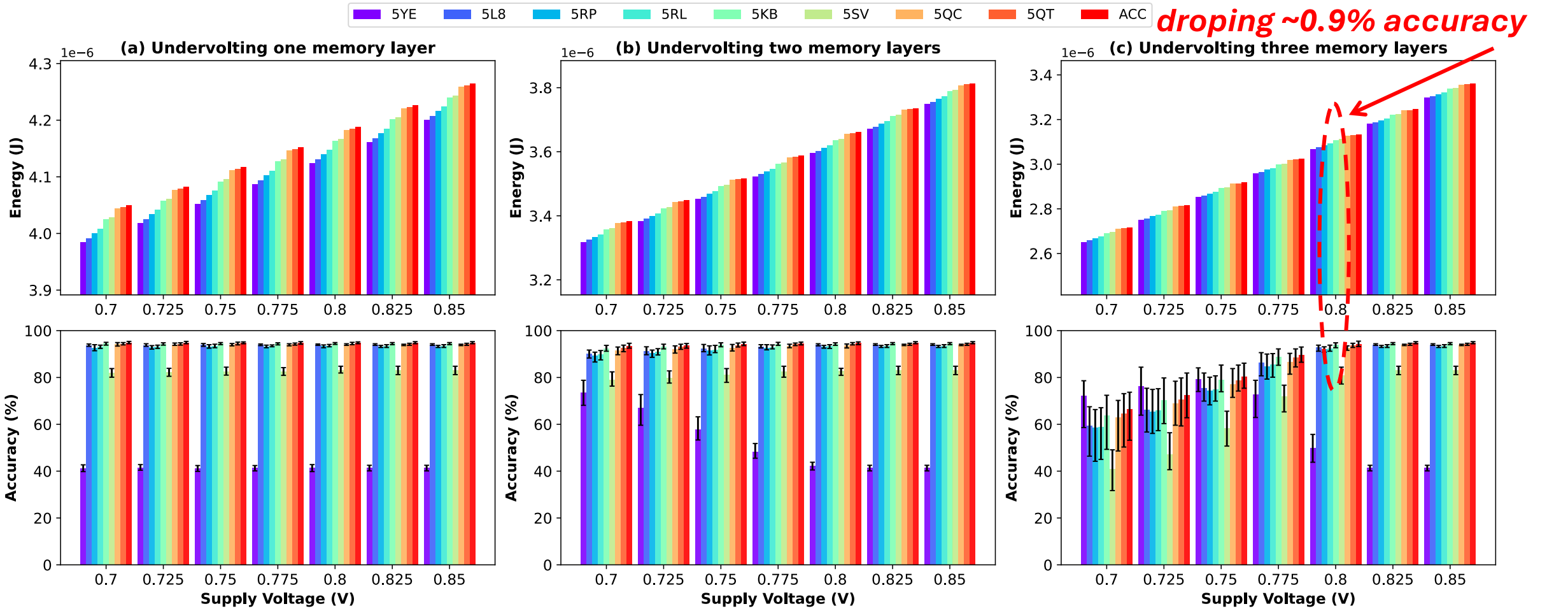
Acc. decreases 0.7%
- ***~24.8% less energy***
- ***~9.9% smaller area***

Acc. decreases 1.5%
- ***~20.9% less energy***
- ***~17.3% smaller area***

Evaluation: Result (2/4)

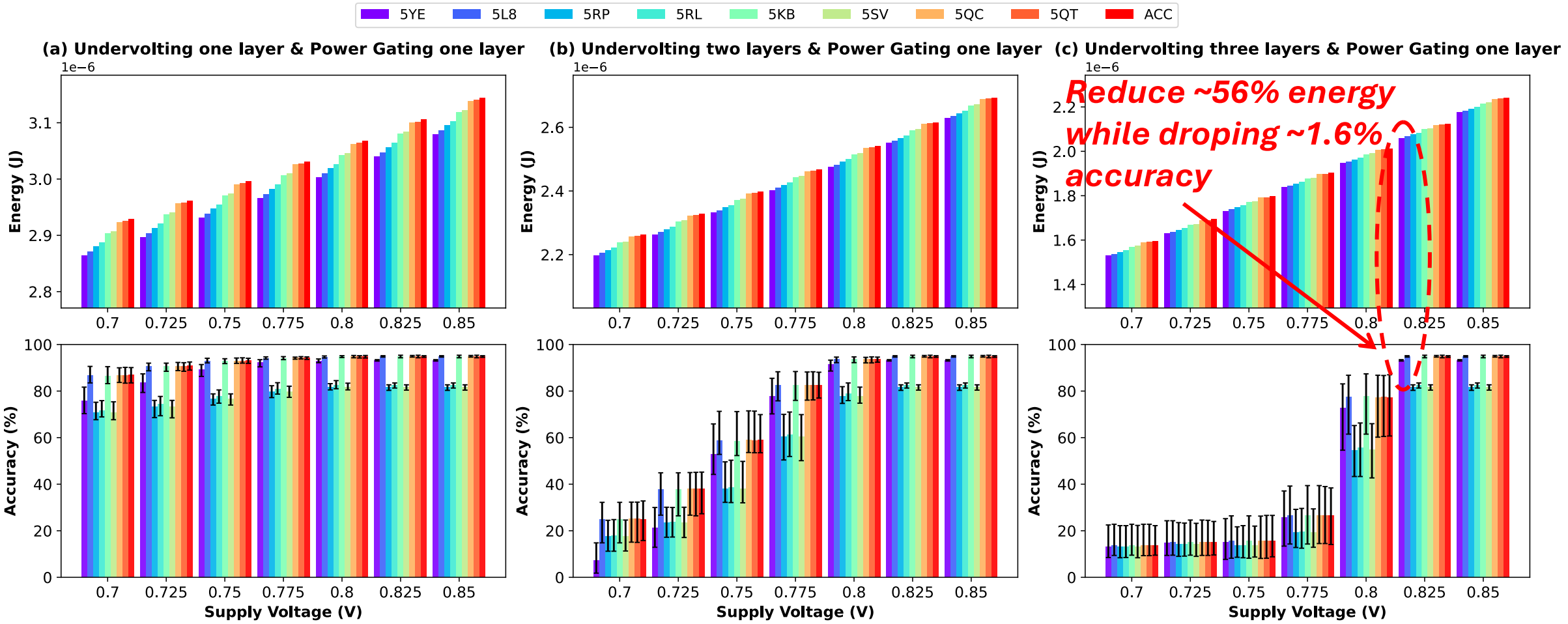
- The transformation with the **Undervolting** technique

Reduce ~31% energy while droping ~0.9% accuracy



Evaluation: Result (3/4)

- The transformation with the **Undervolting & Power-gating** techniques



Evaluation: Result (4/4)

- Comparison results between our hardware and prior works for MNIST

Model Name	Acc.(%)	Arch.	Tech.	Energy per SOP (pJ)	Energy per SOP (pJ) (in 14nm)
TrueNorth [2]	91.94	2D	28nm	26 (0.775V)	4.902
Loihi [3]	96	2D	14nm FinFET	23.6 (0.75V)	23.6
ODIN [4]	84.5	2D	28nm FD-SOI	8.4	1.078
NASH [5]	79.4	3D	45nm	11.3 (1.1V)	0.648
This work	94.8[1]	3D	45nm	20.33[1]	1.167[1]
	93.9[2]			13.28[2]	0.762 [2]
	93.2[3]			8.976[3]	0.515[3]
	77.6[4]			8.374[4]	0.48[4]

[1] Case 1: $\{m_0, m_1, m_2, m_3\} = \{1.1V, 1.1V, 1.1V, 1.1V\}$ (with ACC)
[2] Case 2: $\{m_0, m_1, m_2, m_3\} = \{1.1V, 0.8V, 0.8V, 0.8V\}$ (with 5KB)
[3] Case 3: $\{m_0, m_1, m_2, m_3\} = \{0.825V, 0.825V, 0.825V, 0V\}$ (with 5YE)
[4] Case 4: $\{m_0, m_1, m_2, m_3\} = \{0.8V, 0.8V, 0.8V, 0V\}$ (with 5L8)

Outline

- Research Introduction
- Methodology
- Evaluation
- Conclusion

Conclusion

- Spiking Neural Net. was implemented using approximate neurons and approximate memory.
  - *Neurons* => Approximate Adders
  - *Memory* => **Undervolting** and **Power Gating**
- The approximate neurons reduced energy consumption by up to **24.8%** per neuron with a **0.7%** accuracy loss.
- Combining approximate neurons and approximate memory reduced energy consumption by **55.9%** with a **1.6%** accuracy loss.
- Future work is to optimize the approximate combinations of neurons and memory for energy-optimal Spiking Neural Net.

References

[1] V. Mrazek, et. al., “Evoapprox8b: Library of approximate adders and multipliers for circuit design and benchmarking of approximation methods,” in Design, Automation & Test in Europe Conference & Exhibition (DATE), 2017. IEEE, 2017, pp. 258–261.

[2] F. Akopyan, et al., “Truenorth: Design and tool flow of a 65 mW 1 million neuron programmable neurosynaptic chip,” IEEE Transactions on computer-aided design of integrated circuits and systems, vol. 34, no. 10, pp. 1537–1557, 2015.

[3] M. Davies, et al., “Loihi: A neuromorphic manycore processor with on-chip learning,” IEEE Micro, vol. 38, no. 1, pp. 82–99, 2018.

[4] C. Frenkel, et al., “A 0.086-mm^2 12.7pJ/SOP 64k-synapse 256-neuron online-learning digital spiking neuromorphic processor in 28nm CMOS,” IEEE Transactions on biomedical circuits and systems, vol. 13, no. 1, pp. 145–158, 2018.

[5] O. M. Ikechukwu, et al., “On the design of a fault-tolerant scalable three dimensional NoC-based digital neuromorphic system with on-chip learning,” IEEE Access, vol. 9, pp. 64331– 64 345, 2021.

Thank you for your attention!