The University of Aizu

Research Progress Seminar

Power-aware 3D-stacking-memory Neuromorphic Architecture with Layer-wise Voltage Scaling

Ngo-Doanh NGUYEN - m5262108

2023-05-26

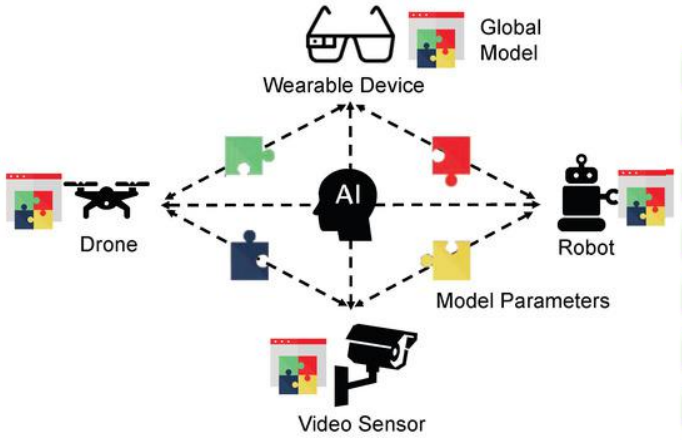
Agenda

1. Motivation
2. Approach & Methodology
3. Proposal Hardware Architecture
4. Results
5. Conclusion

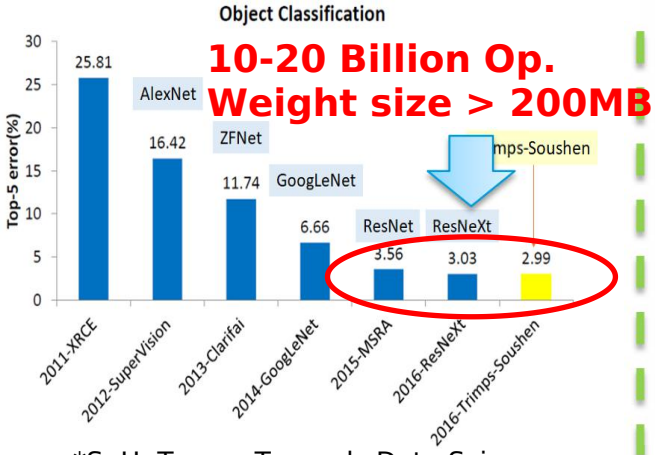
Agenda

1. Motivation
2. Approach & Methodology
3. Proposal Hardware Architecture
4. Results
5. Conclusion

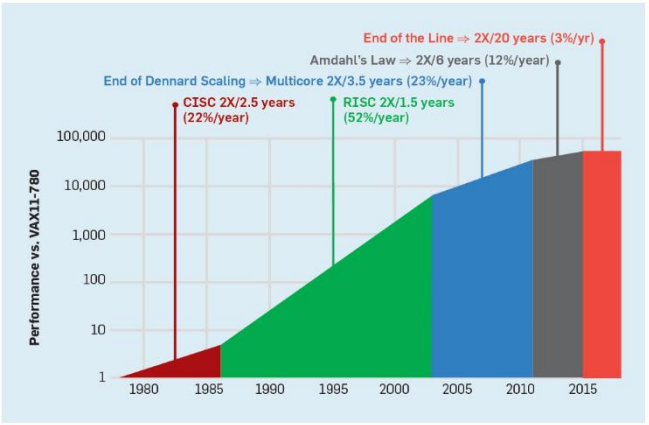
Motivation



*Y. Du. Decentralized Smart IoT. Encyclopedia



*S. H. Tsang. Towards Data Science



*J.Hennessy, D. Patterson 2019 CACM

Computational Power for Edge Devices

AI Enabled Devices

- *Improve Data Transfer Efficiency*
  - *+ Reduce Latency*
  - *+ Reduce Power*

High Complexity for Edge Devices

Spiking Neural Net.

- *Lightweight Inference*
- *Reduce Power Consumption*
- *Reduce Memory Footprint*
- *Reduce Hardware Area*

End of Moore's Law

3-D Stacking Arch.

- *Reduce Latency*
- *Reduce Power Consumption*
- *Reduce Hardware Area*

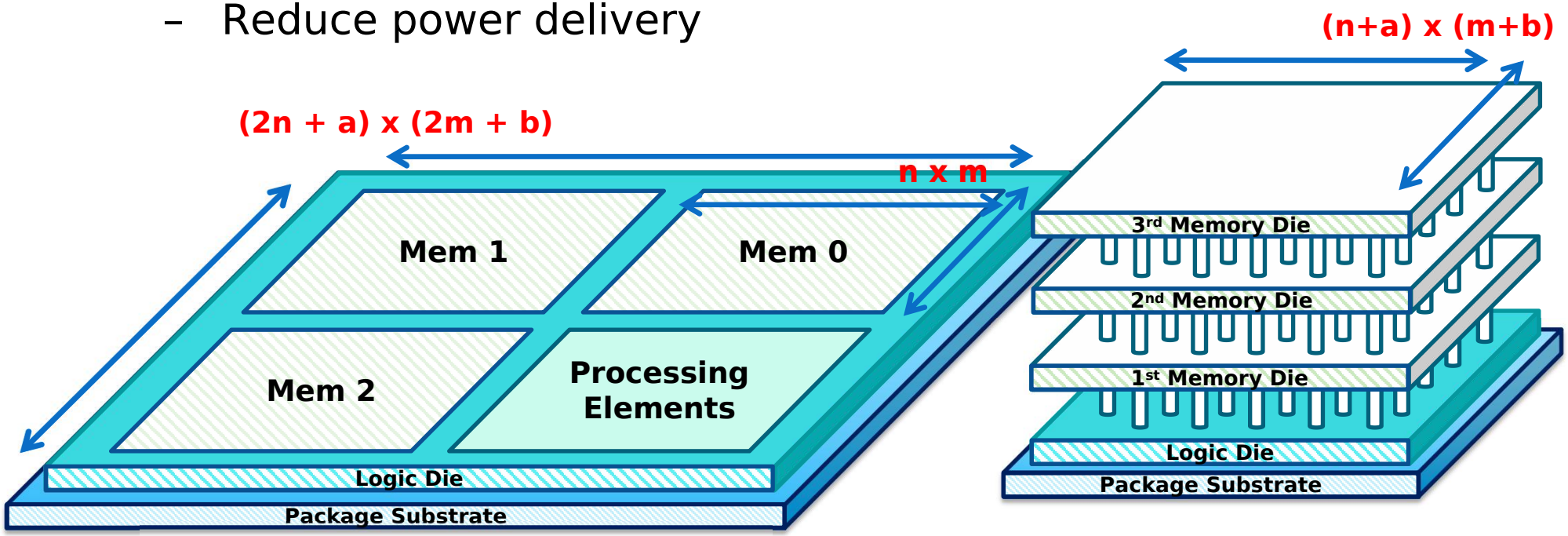
=> Low-power Spiking Neural Network with 3D-stacking-memory

Agenda

1. Motivation
2. Approach
3. Proposal Hardware Architecture
4. Results
5. Conclusion

Approaches (1)

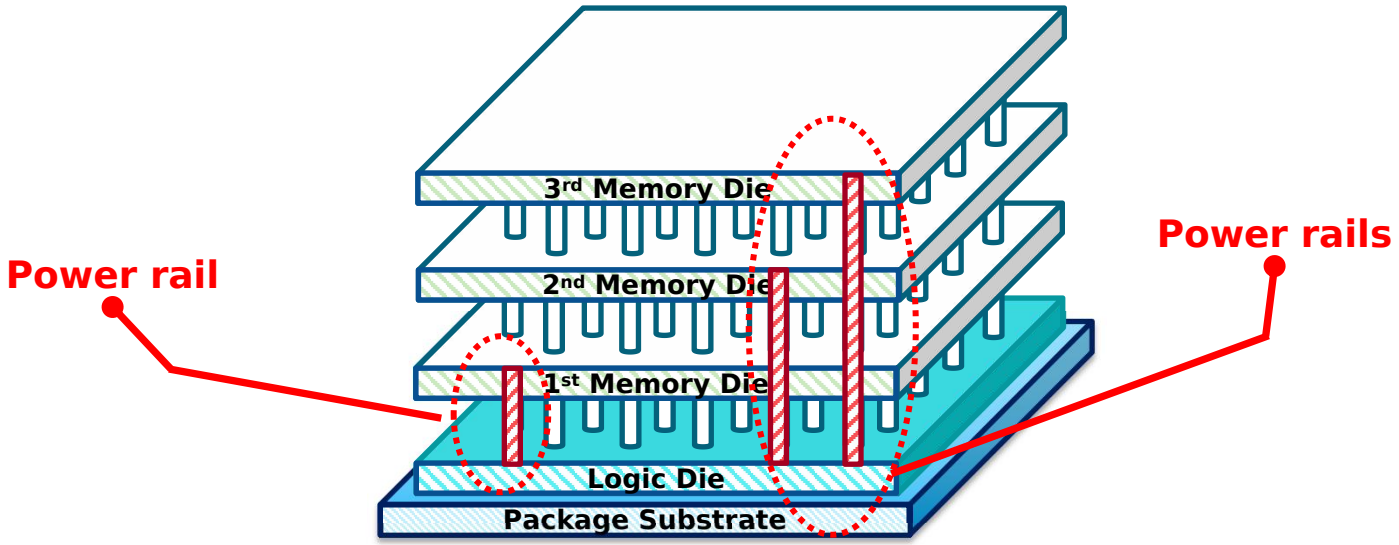
- From 2D Architecture to 3D Architecture
  - Reduce hardware footprint
  - Shorten data movement
  - Reduce power delivery



2D hardware architecture **3D hardware architecture**

Approaches (2)

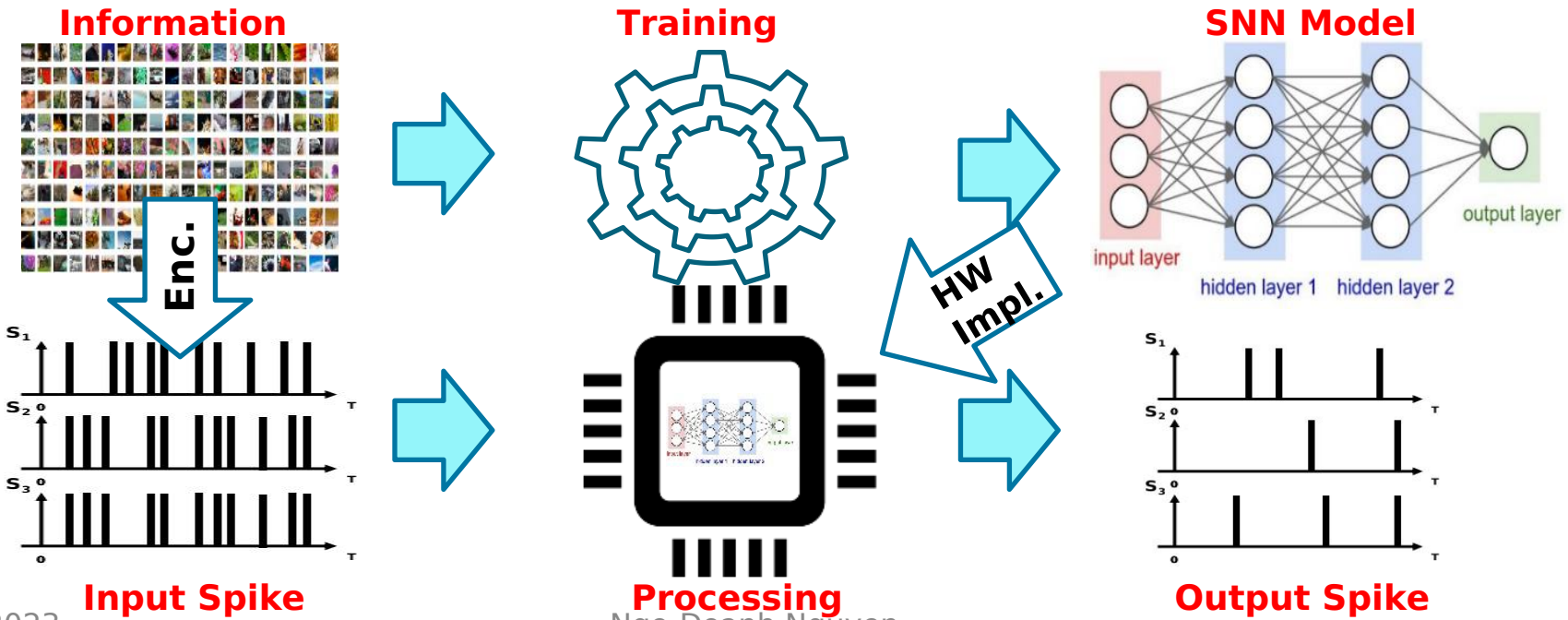
- Each layer in 3D architecture can have isolated power rails
  - Power-gating, voltage-scaling differently for each layer
    - Reduce supply voltage for low-priority layer or power-gate it
    - Maintain supply voltage for high-priority layer



Power management for each layer

Approaches (3)

- Spike Neural Network is spatial and temporal sparse
  - Lightweight inference
  - Low power delivery
- SNN has noise resilience
  - Against the affection of power-gate & voltage-scaling

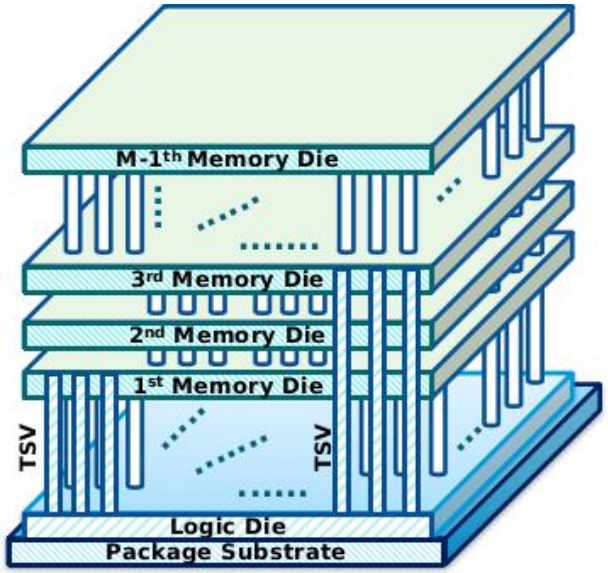


Agenda

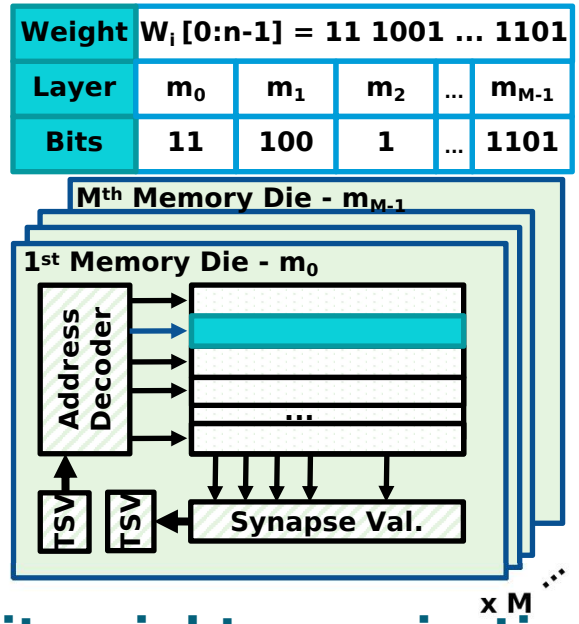
1. Motivation
2. Approach & Methodology
3. Proposal Hardware Architecture
4. Results
5. Conclusion

Proposal Hardware Architecture

- 3D Neuromorphic Computing Core (3D-NCC)
  - Split memory into subsets placed in multiple layers
  - Power-gate & under-voltage are applied separately to each memory layer **=> Lower power consumption**
  - In-situ dynamic quantization for synaptic weights **=> Maintain accuracy**



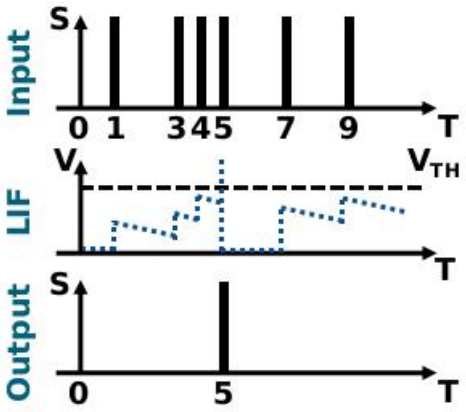
3D-NCC Architecture

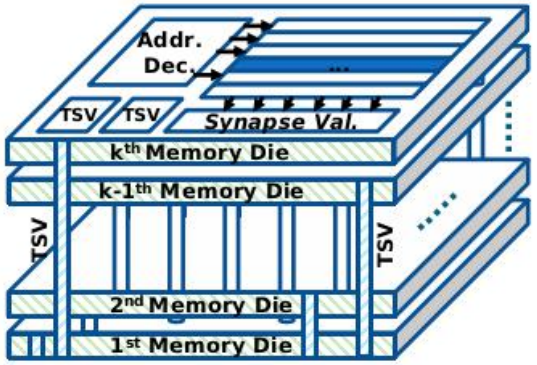


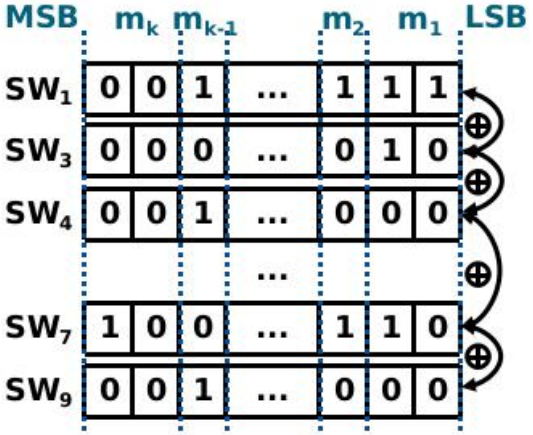
Bit weight organization

Weight operation (1)

- Operation of 3D-NCC under normal condition

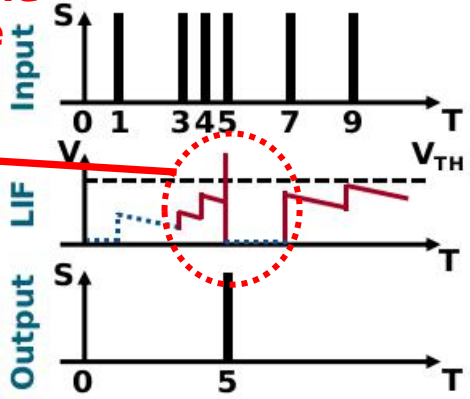




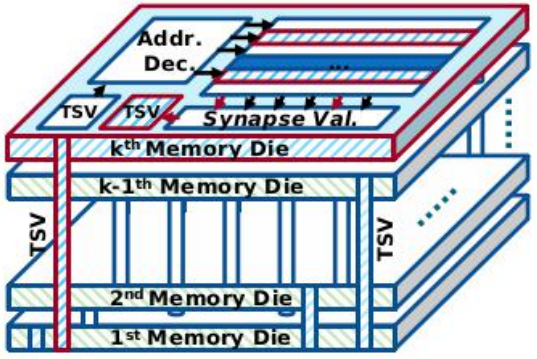


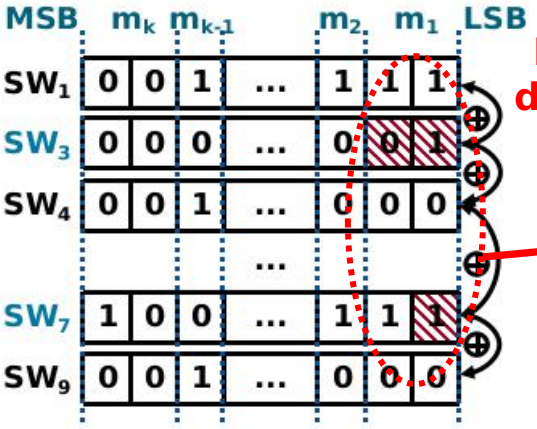
- Operation of 3D-NCC with under-volting top layer

Output stays the same



: Undervolted Defect : Normal

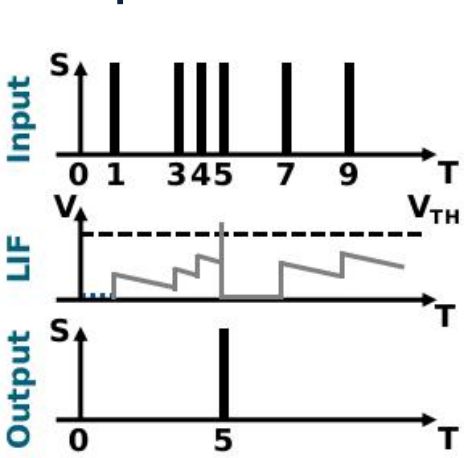


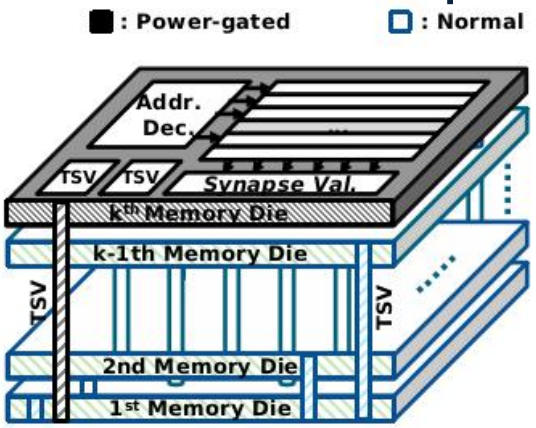


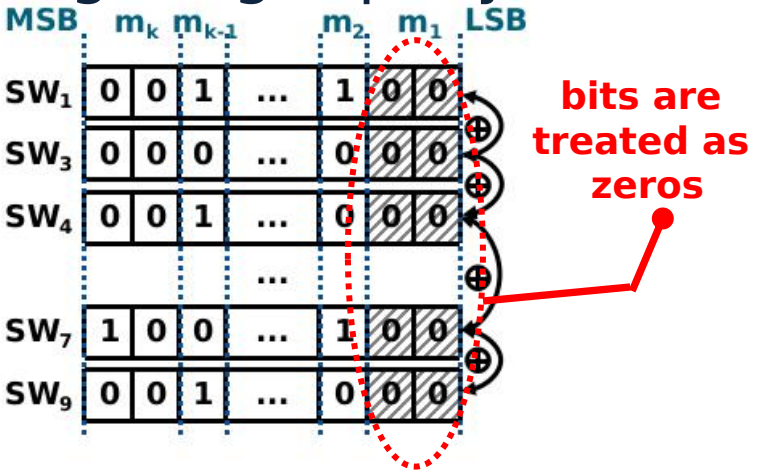
Flipped bits due to under-volting

Weight operation (2)

- Operation of 3D-NCC under power-gating top layer

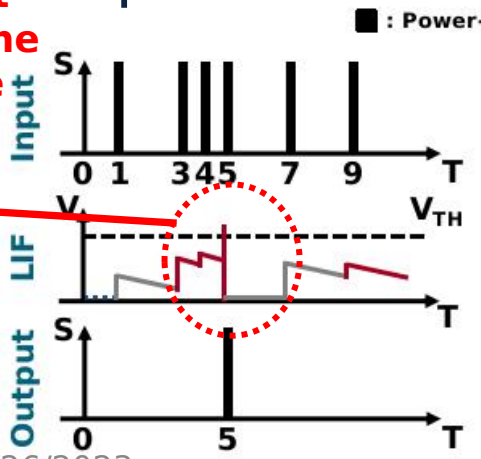


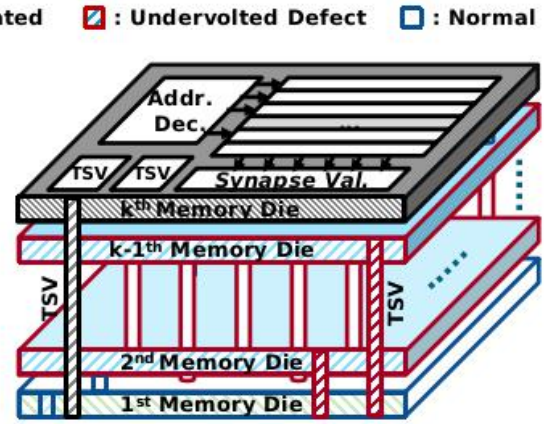


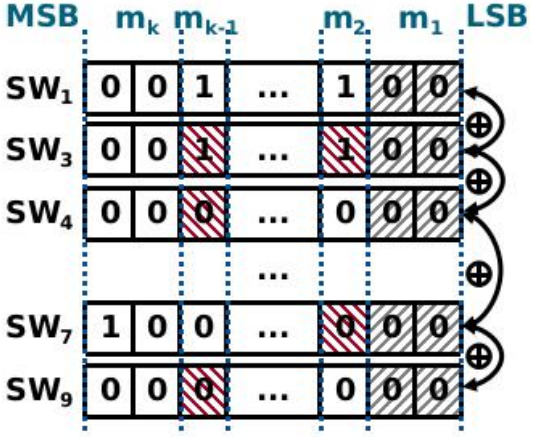


- Operation of 3D-NCC with both PG & UV

Output stays the same





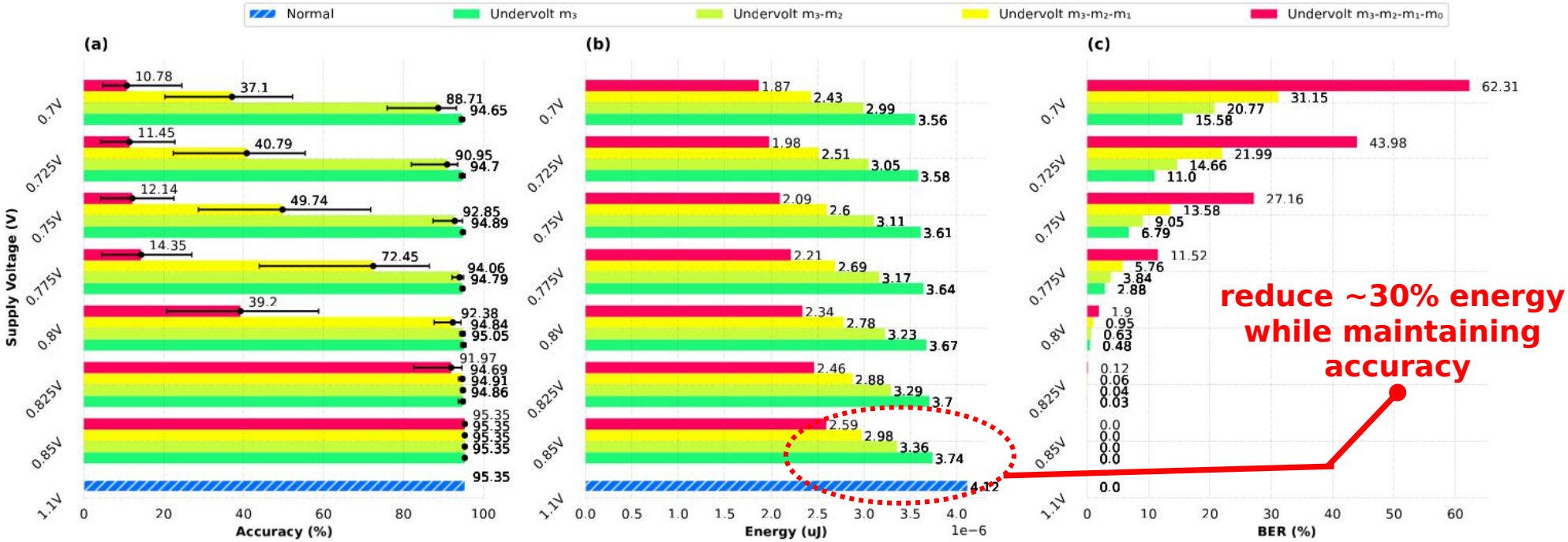


Agenda

1. Motivation
2. Approach & Methodology
3. Proposal Hardware Architecture
4. Results
5. Conclusion

Accuracy vs. Power (1)

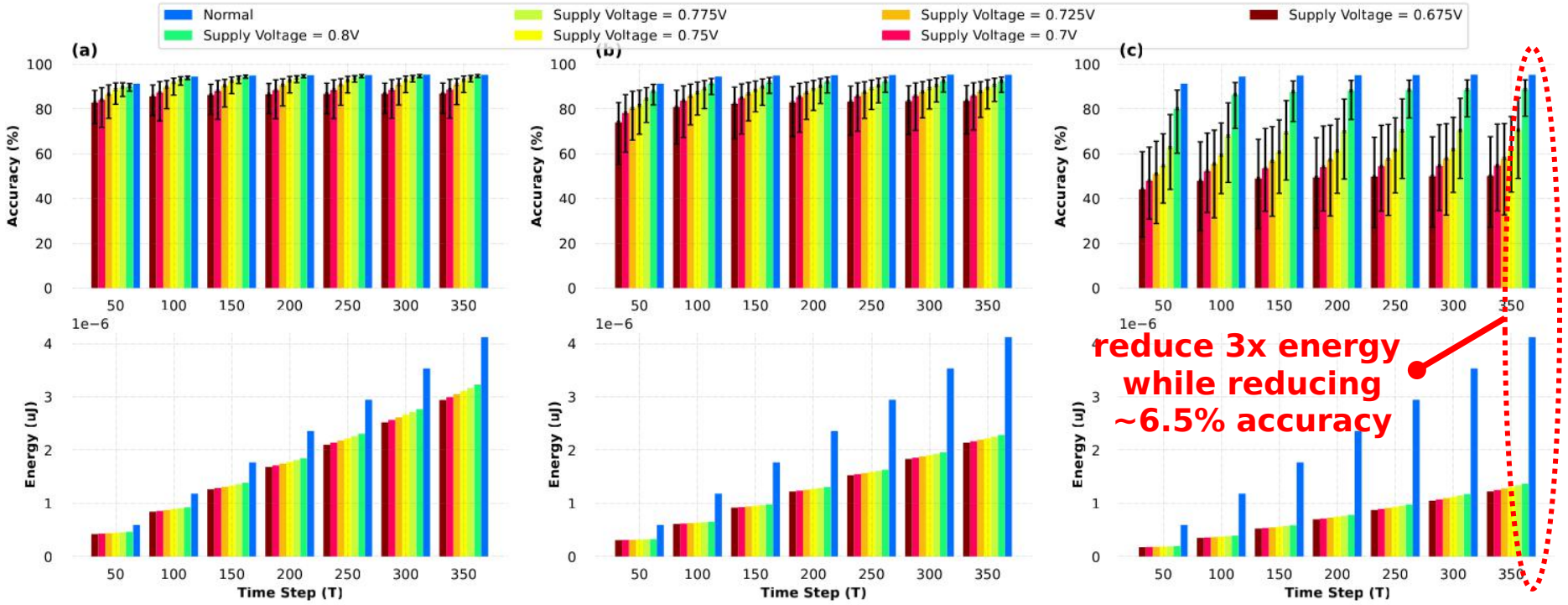
- Dataset: MNIST; SNN config. = 784:48:10 (1 3D-NCC)
- Power extraction with PrimeTime Synopsys
- Undervolt each layer (one-by-one) with the same volt.



Accuracy per Volt. **Energy per Volt.** **Bit Error Rate per Volt.**

Accuracy vs. Power (2)

- Using both power-gating or under-volting for each layer



UV two top layers

PG top layer & UV second top layer

PG two top layers & UV two bottom layers

Hardware Complexity of 3D-NCC

- SNN config. = 784:48:10 (1 3D-NCC)
- Lib: NANGATE 45nm + OpenRAM + FreePDK 3D45 TSV
- Synthesize with Synopsis Design Compiler

Technology		45*nm*
Frequency		100*MHz*
# LIF		48 LIFs
# Stacking Memory		4 layers
# bit of Synaptic Wegihts		8 bits
Bit Configuration in Memory Layer		2-2-2-2
Gate Count	Total (3D-NCC)	809.98*KGEs*
	Memory Blocks	791.76*KGEs*
	Crossbar & Address Decoder	9.68*KGEs*
	LIFs	8.52*KGEs*

Hardware complexity of 1 3D-NCC

Comparison

- Pick two random combinations to compare results

Parameters	Seo *et al.* [56]	Kim *et al.* [57]	TrueNorth [6]	Loihi [1]	ODIN [46]	Ikechukwu *et al.* [58]	Karimi *et al.* [59]	**3D-NCC** (This work)		
								Nom. Case	Case 1[1]	Case 2[2]
Benchmark	MNIST	MNIST	MNIST	MNIST	MNIST	MNIST	MNIST	MNIST		
Accuracy (%)	77.2	84.5	91.94	96	84	79.4	99.2	95.35	94.84	88.77
Neuron Model	LIF	IF	IF	DenMem	LIF & Izhikevicz	LIF	LIF	LIF		
Synaptic Weight Storage	1-bit SRAM	4, 5, 14-bit SRAM	1-bit SRAM	1-to-9-bit SRAM	4-bit SRAM	8-bit SRAM	CTT twin-cell	8-bit SRAM		
Interconnect	2D	2D	2D	2D	2D	3D	2D	3D		
Implementation	Digital	Digital	Digital	Digital	Digital	Digital	Mix-signal	Digital		
Learning Rule	On-chip STDP	Stochastic Gradient Descent	Un-supervised	On-chip STDP	On-chip Stochastic SDSP	On-chip STDP	Off-chip	Off-chip		
Technology	45nm SOI	65nm	28nm	14nm FinFET	28nm FD-SOI	45nm	22nm FD-SOI	45nm		
Supply Voltage	0.55-1 V	0.45 V	0.7-1.05V	0.5-1.2 V	0.55-1 V	1.1 V	0.8 V	0.65V - 1.1V		
Energy per SOP (pJ)	N/A	N/A	26 (0.775V)	23.6 (0.75V)	8.4	189.3	8	244.28	191.46	81.16
Energy per SOP (pJ) (in 14nm)	N/A	N/A	4.902	23.6	1.078	10.86	4.32	14.02	10.98	4.65

[1] Case 1: $\{V_{m_0} = 1.1V; V_{m_1} = 1.1V; V_{m_2} = 0.8V; V_{m_3} = 0.8V\}$ (Low-power Mode I)
[2] Case 2: $\{V_{m_0} = 0.825V; V_{m_1} = 0.8V; V_{m_2} = 0V; V_{m_3} = 0V\}$ (Low-power Mode III)

likely same energy with difference in technology

Agenda

1. Motivation
2. Approach & Methodology
3. Proposal Hardware Architecture
4. Results
5. Conclusion

Conclusion

- 3D SNN architecture called 3D-NCC
- Split memory word into subsets placed in separated layer
  - Apply power-gating & voltage-scaling to memory layer(s)
    - Reduce power consumption while maintaining accuracy
  - In-situ dynamic quantization
- UV two top layer reduces **21.62%** power consumption with **0.51%** accuracy loss
- PG two top layer & UV two bottom layers reduce **66.77%** power consumption with **6.58%** accuracy loss

Schedule

- 29/05 - 11/06: Submit the 1st journal paper extended from conference paper
- 05/06 - 18/06: Submit the 2nd journal paper
- 19/06 - 01/09: Start working on FPGA implementation for SNN
- 19/06 - 01/08: Write draft for conference (MCSoC)

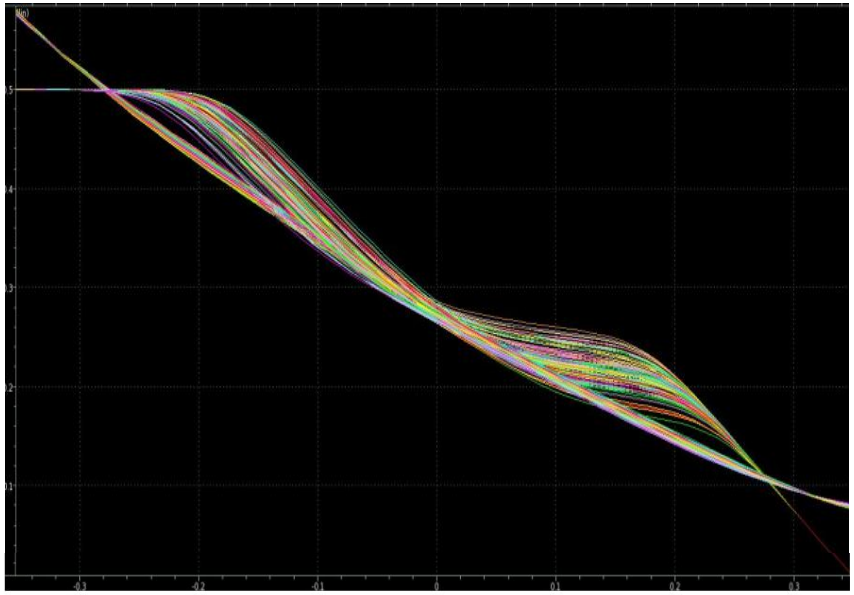
The University of Aizu

Thank you

for your attention.

Signal Noise Margin

- SNM is to get the Bit Error Rate of memory (SNM<0.1)
- 6T SRAM (FreePDK NANGATE 45nm)
- HSPICE simulation + mathematical computations
- Monte Carlo simulation



Signal Noise Margin of 6T SRAM