The University of Aizu

Research Progress Report Seminar

Spiking Neural Network with 3-D IC-based Stacking Memory

Ngo-Doanh NGUYEN - m5262108

Supervised by Prof. DANG Nam Khanh

2023-12-01

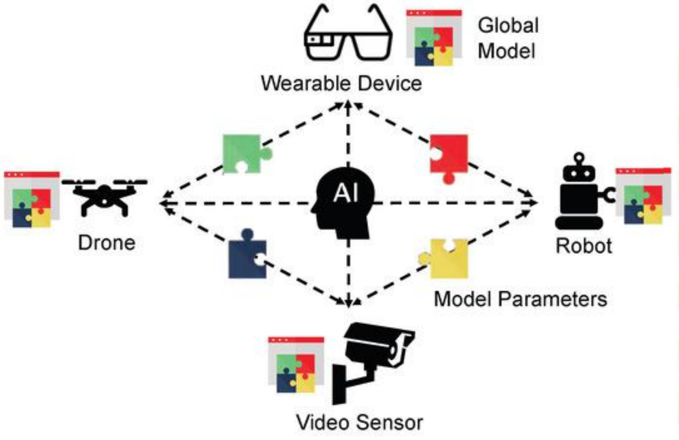
Agenda

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

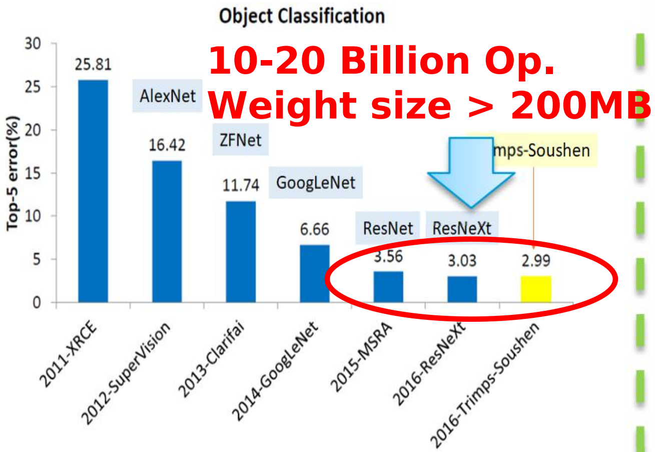
Agenda

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

Motivation

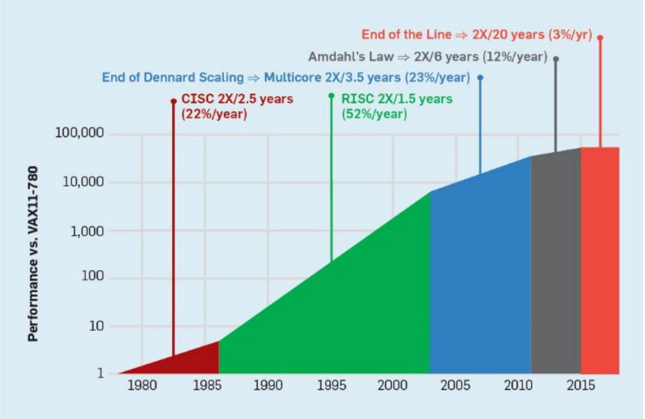


*Y. Du. Decentralized Smart IoT. Encyclopedia



10-20 Billion Op.
Weight size > 200MB

*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 Footprint

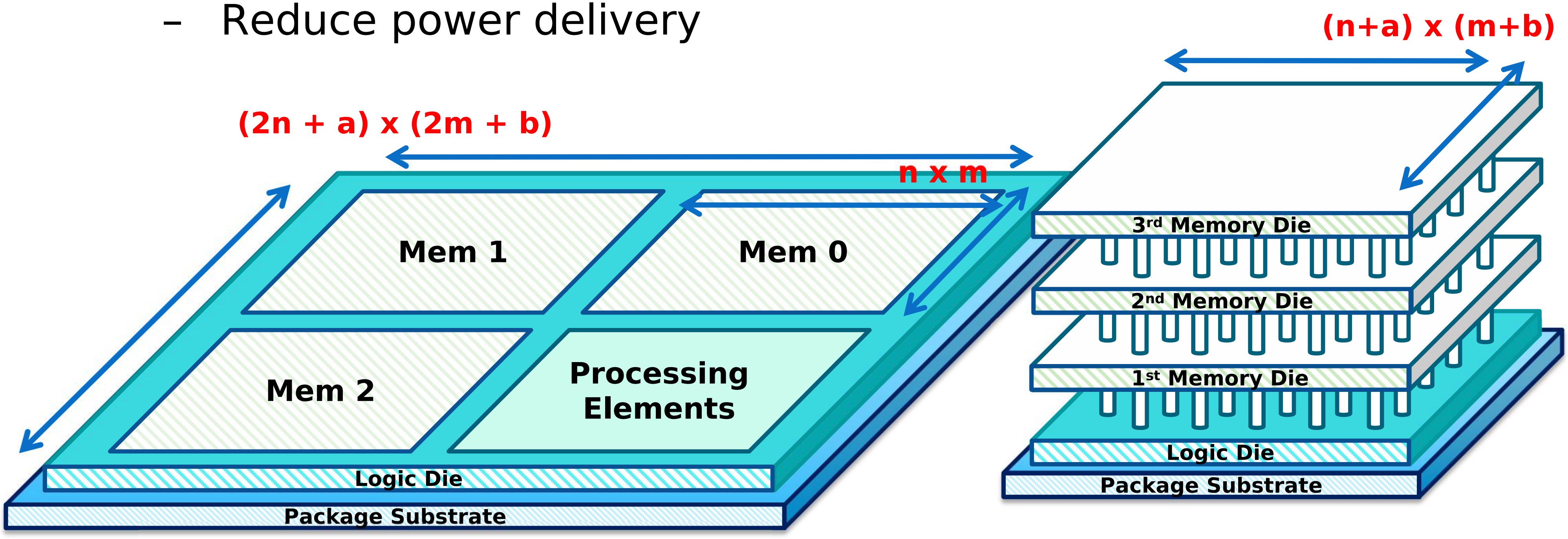
=> Low-power Spiking Neural Network with 3D-stacking-memory

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

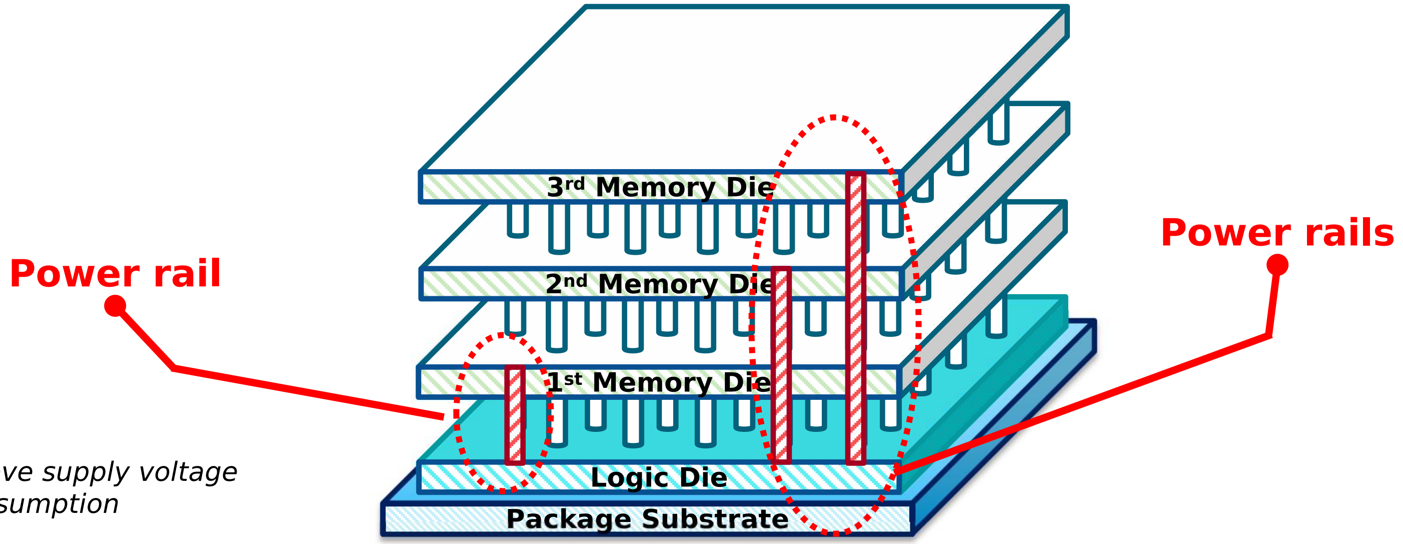


2D hardware architecture

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



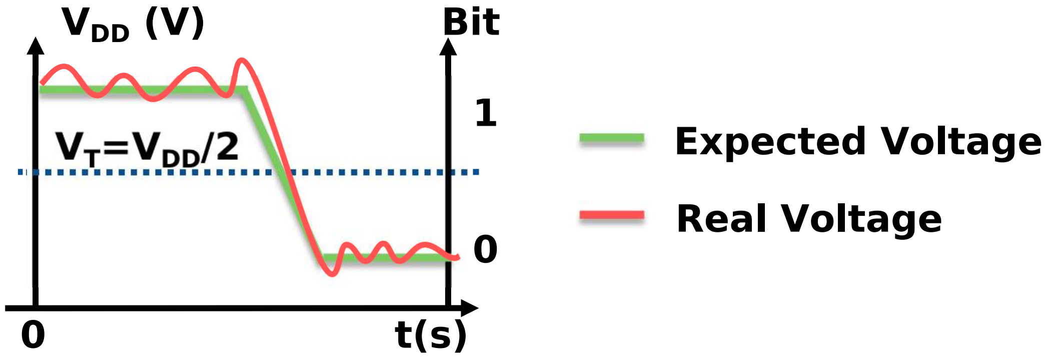
Power-gate = Remove supply voltage to reduce power consumption

***Voltage-scaling = Reduce supply voltage to reduce power consumption*

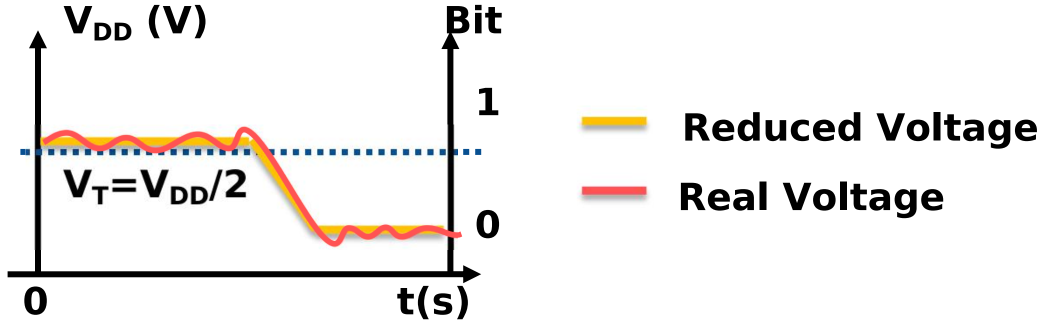
Power management for each layer

Recap: Supply Voltage in Digital Circuits

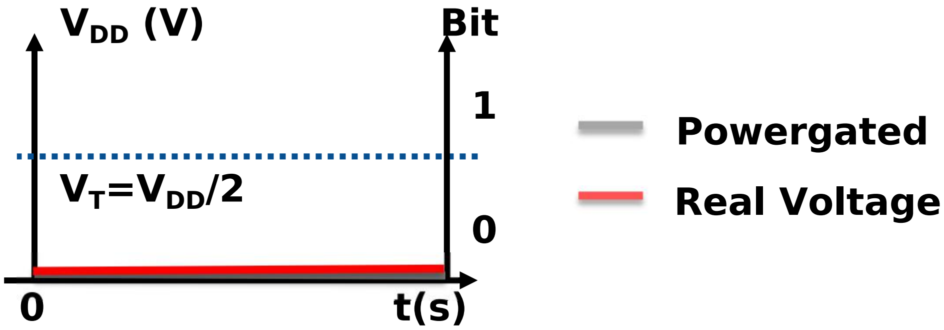
- Normal operation



- Undervolting operation

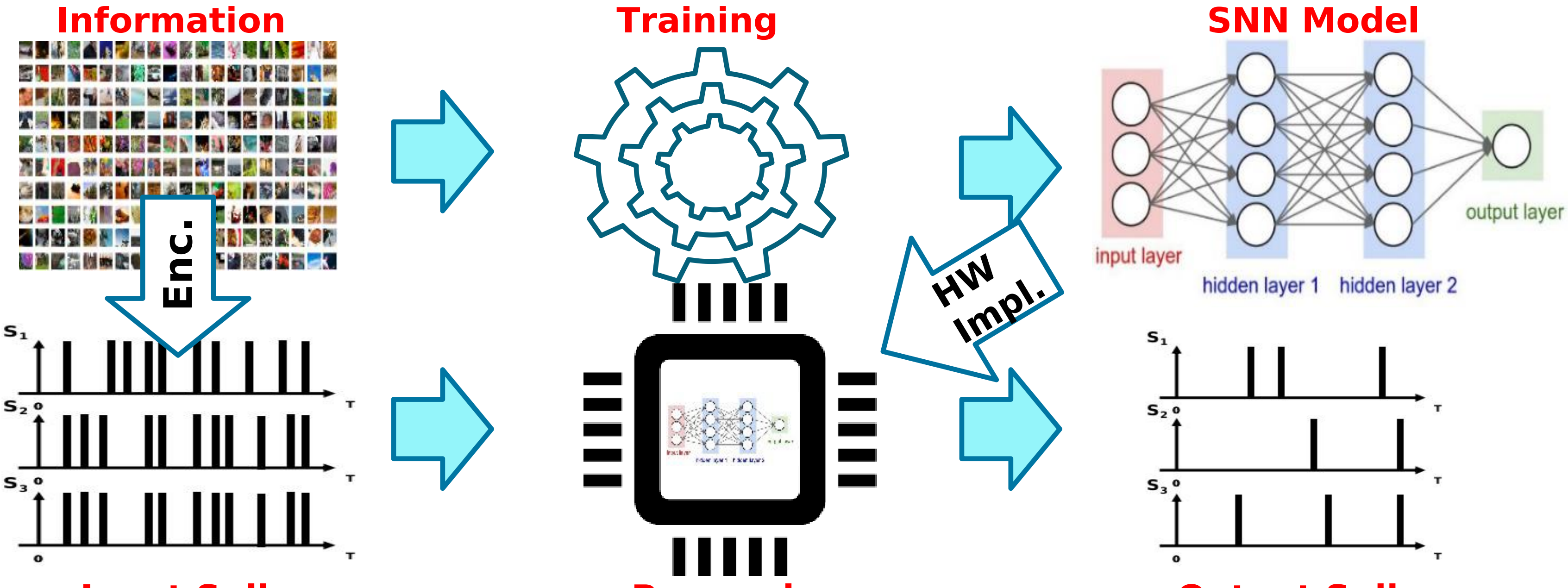


- Powergating operation



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



Information

Training

SNN Model

Enc.

HW Impl.

Input Spike

Processing

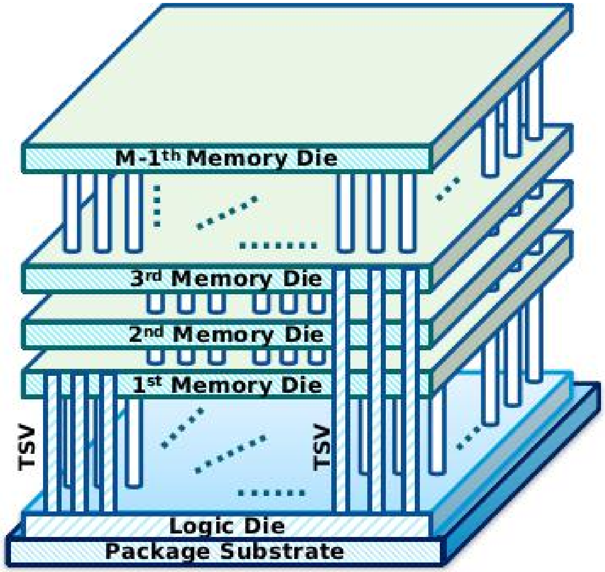
Output Spike

Agenda

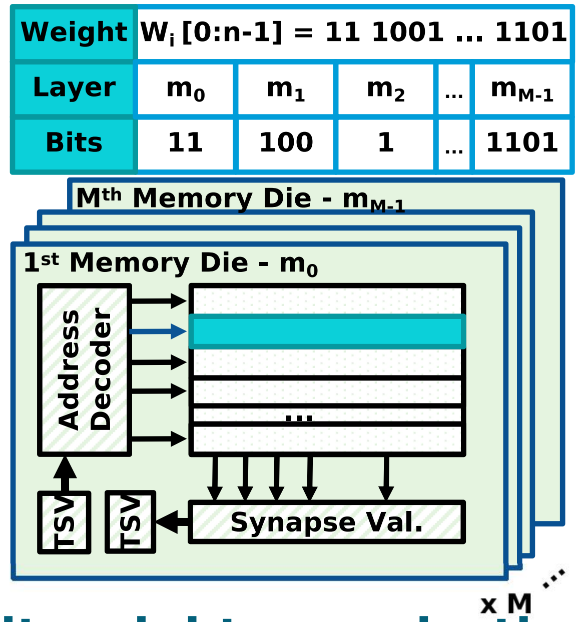
1. Motivation
2. Approach & Methodology
3. Proposal Hardware Architecture
4. Results
5. Conclusion

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

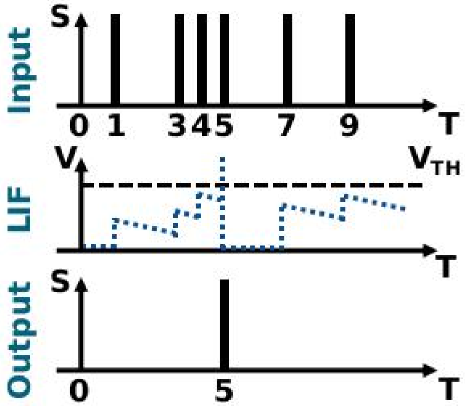


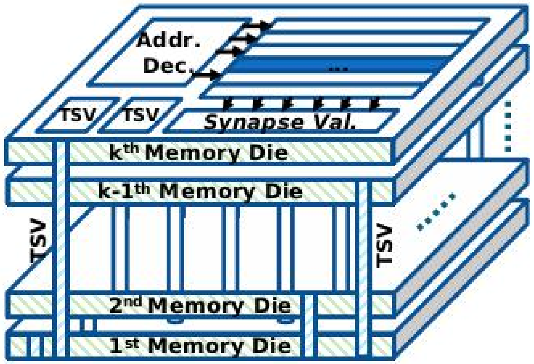
Weight	W_i [0:n-1] = 11 1001 … 1101				
Layer	m_0	m_1	m_2	…	m_{M-1}
Bits	11	100	1	…	1101

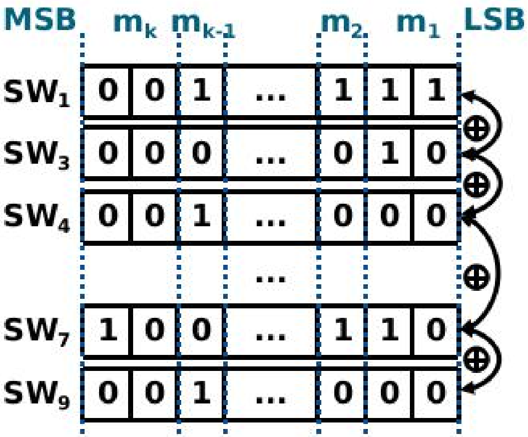
Bit weight organization

Weight operation (1)

- Operation of 3D-NCC under normal condition



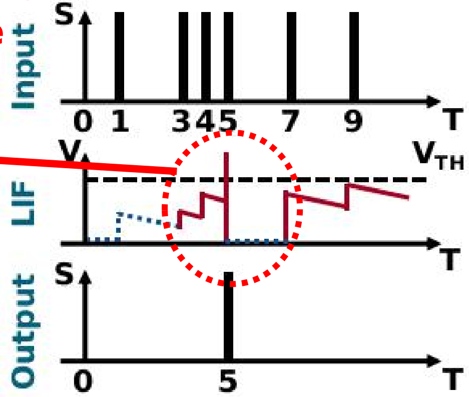




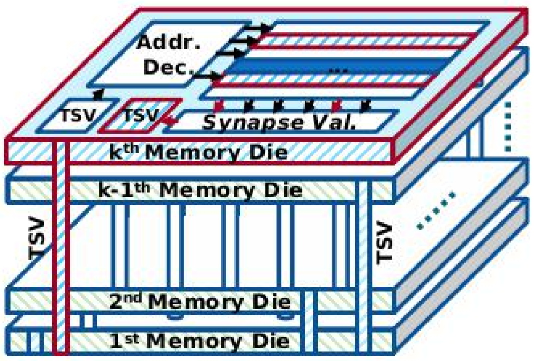
	MSB m_k		m_{k-1}		m_2	m_1	LSB
SW_1	0	0	1	...	1	1	1
SW_3	0	0	0	...	0	1	0
SW_4	0	0	1	...	0	0	0
				...			
SW_7	1	0	0	...	1	1	0
SW_9	0	0	1	...	0	0	0

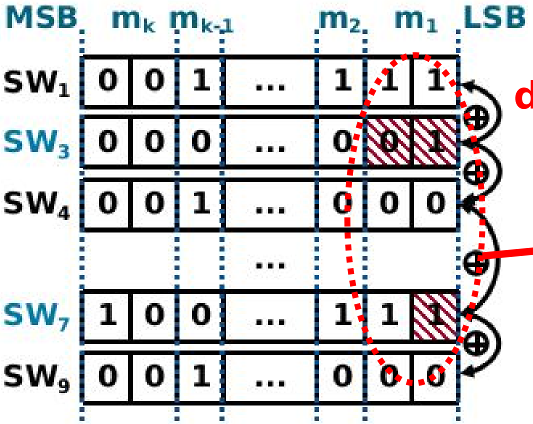
- Operation of 3D-NCC with under-volting top layer

Output stays the same



▨ : Undervolted Defect □ : Normal



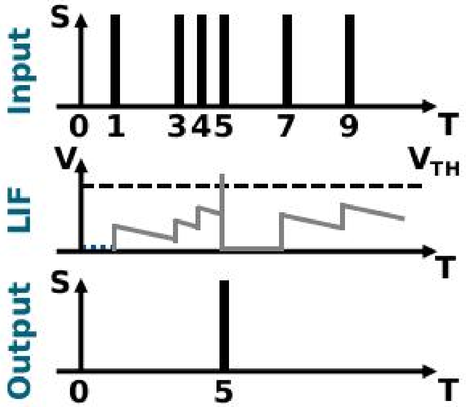


	MSB m_k		m_{k-1}		m_2	m_1	LSB
SW_1	0	0	1	...	1	1	1
SW_3	0	0	0	...	0	0	1
SW_4	0	0	1	...	0	0	0
				...			
SW_7	1	0	0	...	1	1	1
SW_9	0	0	1	...	0	0	0

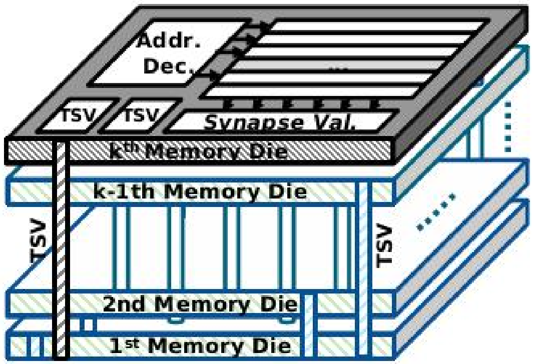
Flipped bits due to under-volting

Weight operation (2)

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



■ : Power-gated □ : Normal

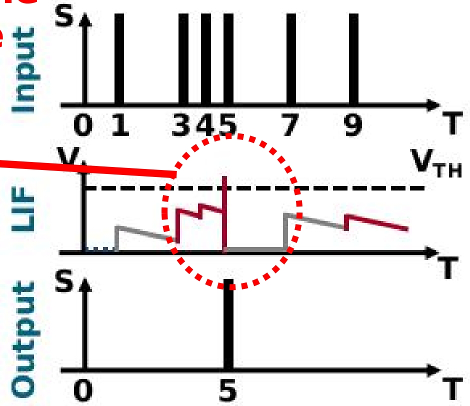


	MSB	m_k	m_{k-1}		m_2	m_1	LSB
SW_1	0	0	1	...	1	▨0	▨0
SW_3	0	0	0	...	0	▨0	▨0
SW_4	0	0	1	...	0	▨0	▨0
				...			
SW_7	1	0	0	...	1	▨0	▨0
SW_9	0	0	1	...	0	▨0	▨0

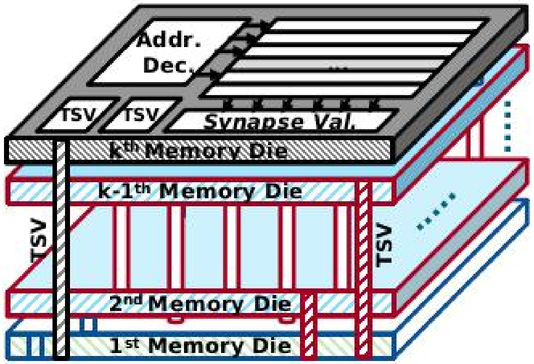
bits are treated as zeros

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

Output stays the same



■ : Power-gated ▨ : Undervolted Defect □ : Normal



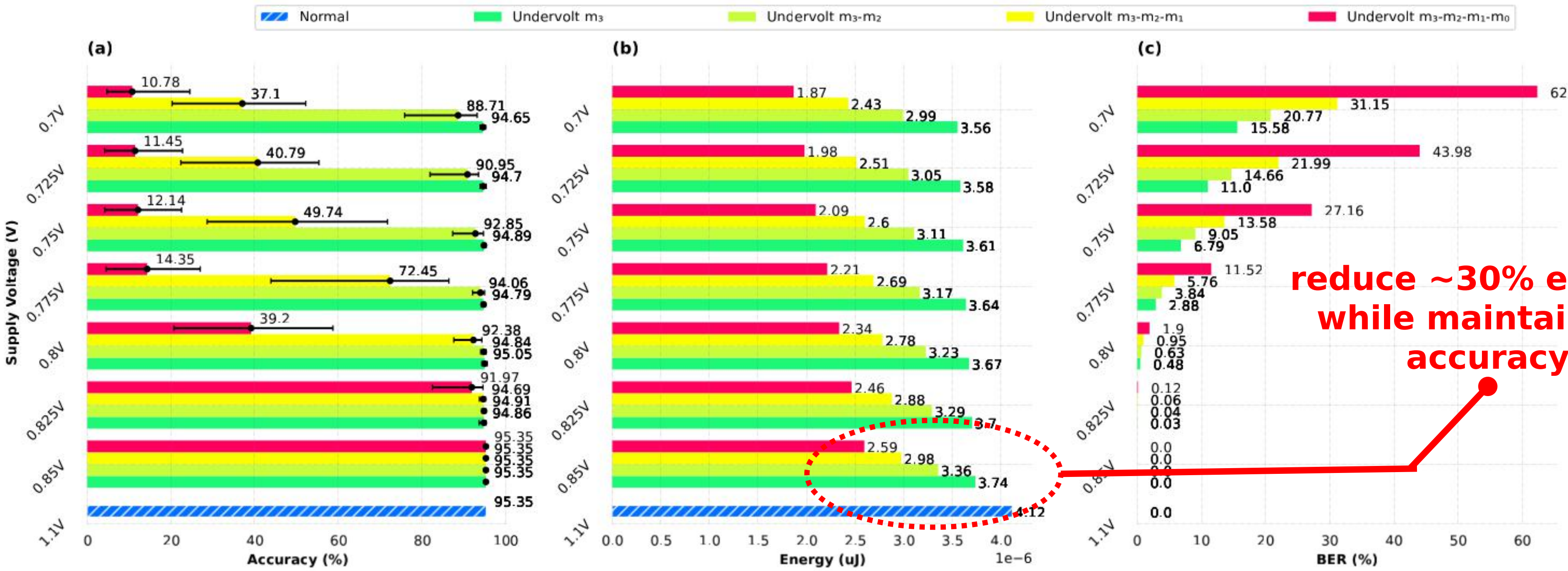
	MSB	m_k	m_{k-1}		m_2	m_1	LSB
SW_1	0	0	1	...	1	▨0	▨0
SW_3	0	0	▨1	...	▨1	▨0	▨0
SW_4	0	0	▨0	...	0	▨0	▨0
				...			
SW_7	1	0	0	...	▨0	▨0	▨0
SW_9	0	0	▨0	...	0	▨0	▨0

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
- Lib: NANGATE 45nm + OpenRAM + FreePDK 3D45 TSV
- Undervolt each layer (one-by-one) with the same volt.

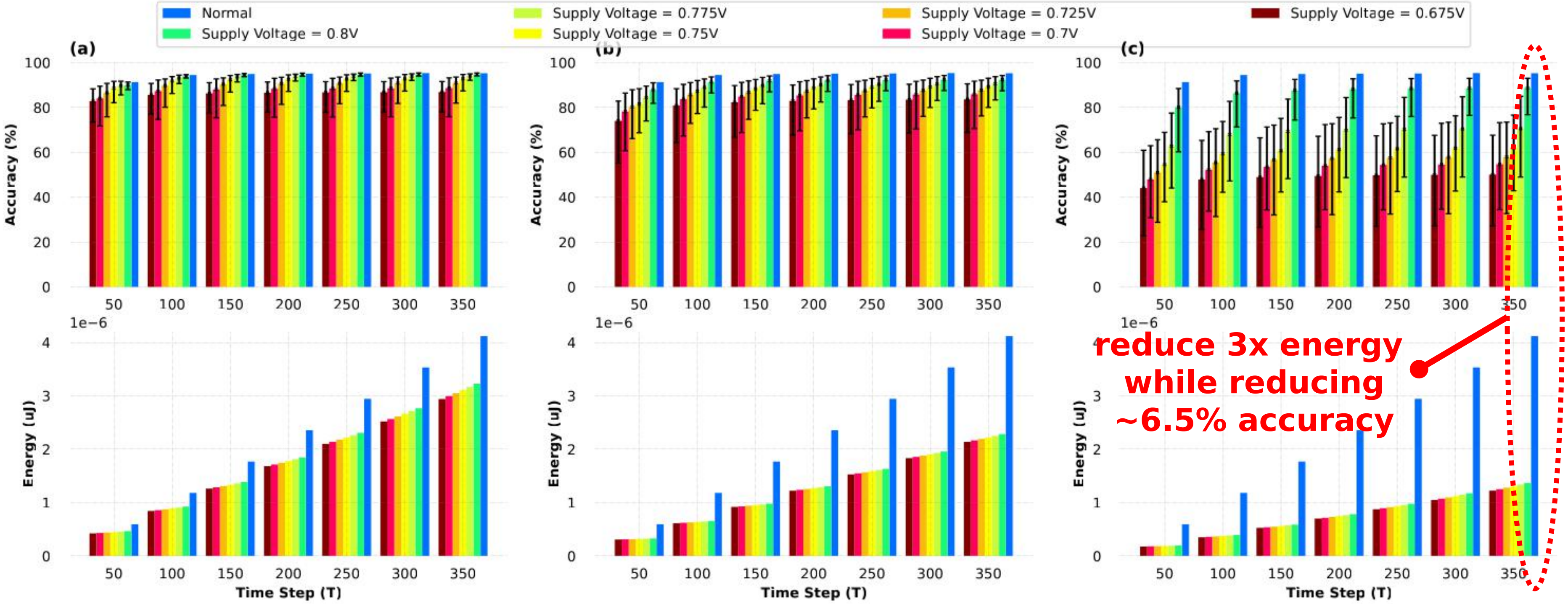


reduce ~30% energy while maintaining accuracy

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



reduce 3x energy while reducing ~6.5% accuracy

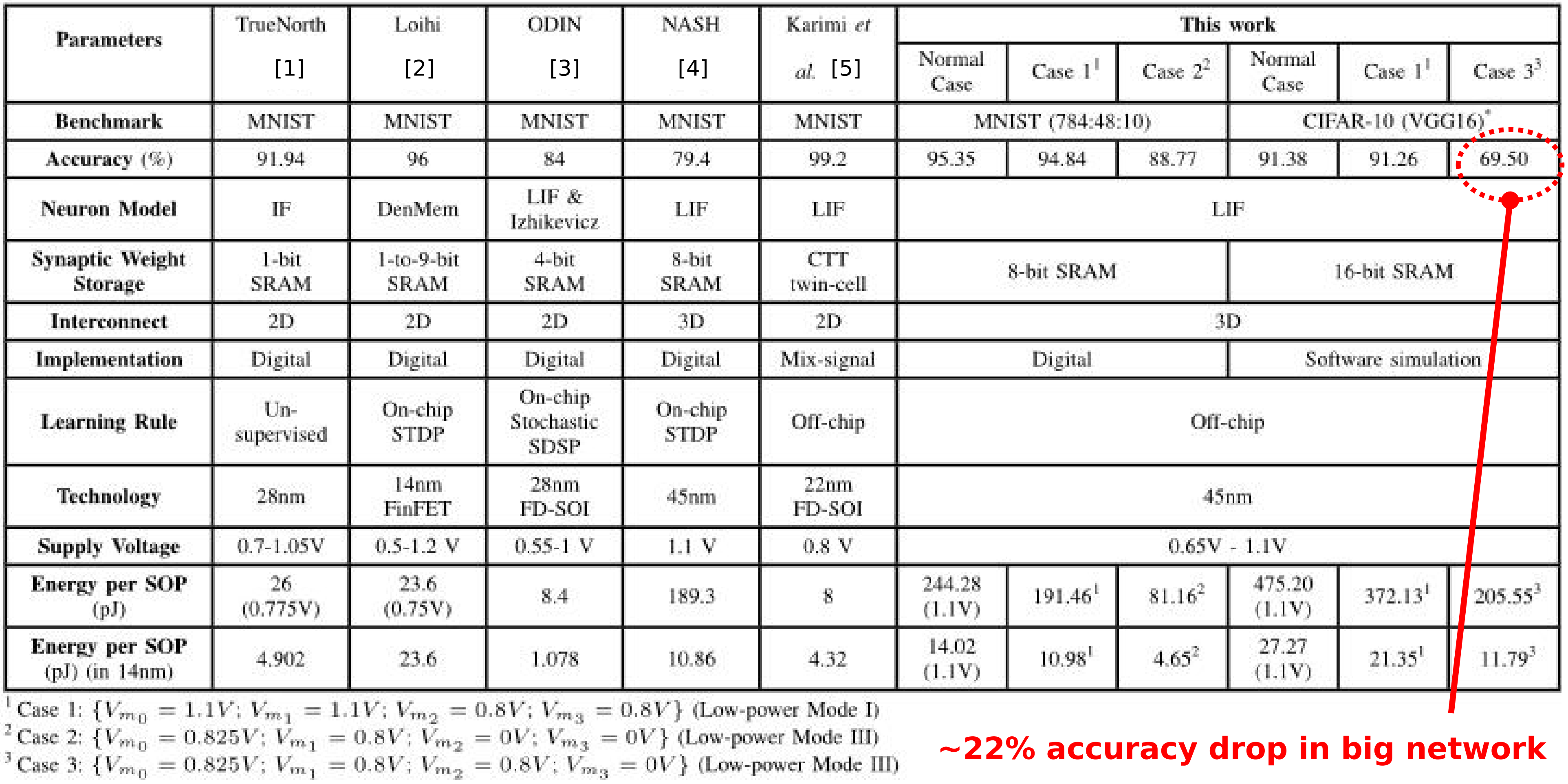
UV two top layers

PG top layer & UV second top layer

PG two top layers & UV two bottom layers

Comparison

- Setup voltage supplies as in case 1, case 2 and case 3



Parameters	TrueNorth [1]	Loihi [2]	ODIN [3]	NASH [4]	Karimi *et al.* [5]	This work					
						Normal Case	Case 1[1]	Case 2[2]	Normal Case	Case 1[1]	Case 3[3]
Benchmark	MNIST	MNIST	MNIST	MNIST	MNIST	MNIST (784:48:10)			CIFAR-10 (VGG16)*		
Accuracy (%)	91.94	96	84	79.4	99.2	95.35	94.84	88.77	91.38	91.26	69.50
Neuron Model	IF	DenMem	LIF & Izhikevicz	LIF	LIF	LIF					
Synaptic Weight Storage	1-bit SRAM	1-to-9-bit SRAM	4-bit SRAM	8-bit SRAM	CTT twin-cell	8-bit SRAM			16-bit SRAM		
Interconnect	2D	2D	2D	3D	2D	3D					
Implementation	Digital	Digital	Digital	Digital	Mix-signal	Digital			Software simulation		
Learning Rule	Un-supervised	On-chip STDP	On-chip Stochastic SDSP	On-chip STDP	Off-chip	Off-chip					
Technology	28nm	14nm FinFET	28nm FD-SOI	45nm	22nm FD-SOI	45nm					
Supply Voltage	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)	26 (0.775V)	23.6 (0.75V)	8.4	189.3	8	244.28 (1.1V)	191.46[1]	81.16[2]	475.20 (1.1V)	372.13[1]	205.55[3]
Energy per SOP (pJ) (in 14nm)	4.902	23.6	1.078	10.86	4.32	14.02 (1.1V)	10.98[1]	4.65[2]	27.27 (1.1V)	21.35[1]	11.79[3]

[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)

[3] Case 3: $\{V_{m_0} = 0.825V;\ V_{m_1} = 0.8V;\ V_{m_2} = 0.8V;\ V_{m_3} = 0V\}$ (Low-power Mode III)

~22% accuracy drop in big network

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

- Prepare & submit the draft of the master's thesis

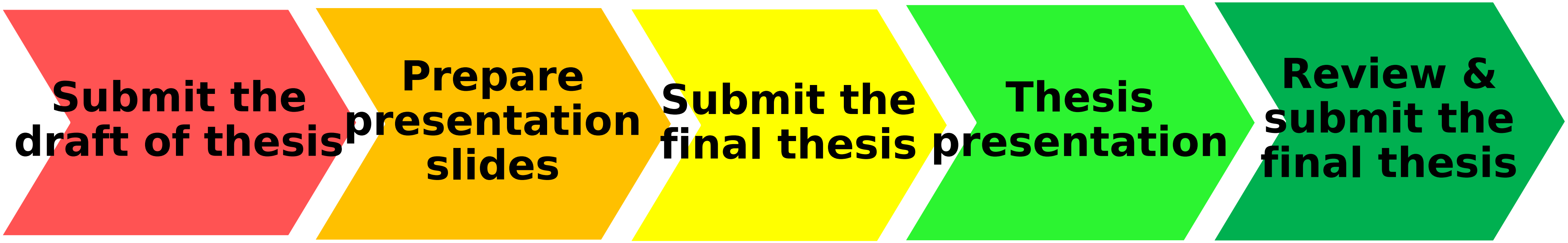
Now - 04.12.23

- Polish the draft of master's thesis and submit the final version

04.12.23 - 26.01.24

- Fix the thesis according the reviews of referees

16.02.24 - 22.02.24



Submit the draft of thesis

Prepare presentation slides

Submit the final thesis

Thesis presentation

Review & submit the final thesis

04.12.23 - 26.12.23

- Prepare the presentation slides based on publication materials

26.01.24 - 16.02.24

- The presentation day (16.02.24)

References

[1] F. Akopyan et al., “TrueNorth: Design and tool flow of a 65 mW 1 million neuron programmable neurosynaptic chip,” IEEE Trans. Comput. Aided Design Integr. Circuits Syst., vol. 34, no. 10, pp. 1537–1557, Oct. 2015.

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

[3] C. Frenkel, C. Frenkel, M. Lefebvre, J.-D. Legat, and D. Bol, “A 0.086-mm2 12.7-pJ/SOP 64k-synapse 256-neuron online-learning digital spiking neuromorphic processor in 28nm CMOS,” IEEE Trans. Biomed. Circuits Syst., vol. 13, no. 1, pp. 145–158, Feb. 2019.

[4] O. M. Ikechukwu, K. N. Dang, and A. B. Abdallah, “On the design of a fault-tolerant scalable three dimensional NoC-based digital neuromorphic system with on-chip learning,” IEEE Access, vol. 9, pp. 64331–64345, 2021.

[5] M. Karimi, A. S. Monir, R. Mohammadrezaee, and B. Vaisband, “CTT-based scalable neuromorphic architecture,” IEEE J. Emerg. Sel. Topics Circuits Syst., vol. 13, no. 1, pp. 96–107, Mar. 2023.

Publications

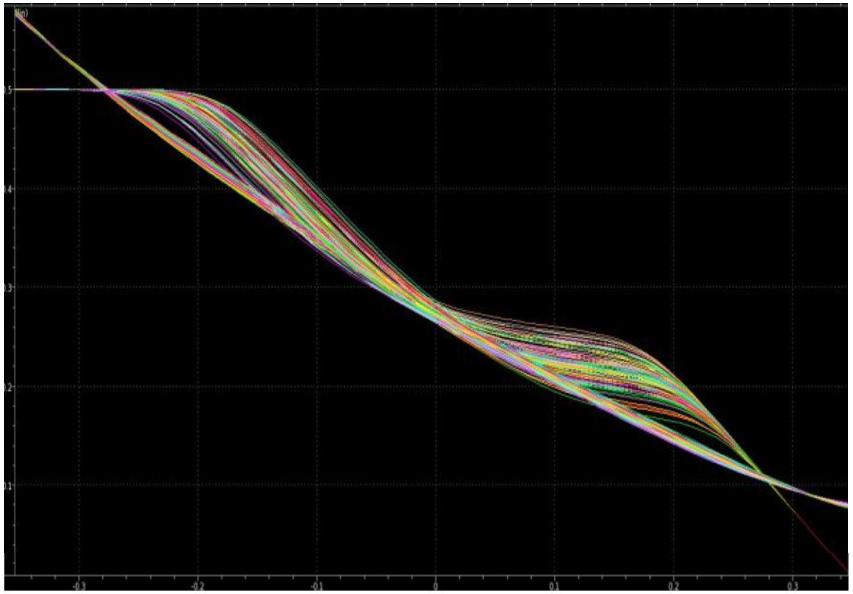
- Journal (Published)
  - N. -D. Nguyen, A. B. Ahmed, A. B. Abdallah and K. N. Dang, "Power-Aware Neuromorphic Architecture With Partial Voltage Scaling 3-D Stacking Synaptic Memory," in IEEE Transactions on Very Large Scale Integration (VLSI) Systems, doi: 10.1109/TVLSI.2023.3318231.
  - N. -D. Nguyen, X. -T. Tran, A. B. Abdallah and K. N. Dang, "An In-Situ Dynamic Quantization With 3D Stacking Synaptic Memory for Power-Aware Neuromorphic Architecture," in IEEE Access, vol. 11, pp. 82377-82389, 2023, doi: 10.1109/ACCESS.2023.3301560.
- Conference: (Accepted)
  - N. -D. Nguyen and K. N. Dang, “A Novel Yield Improvement Approach for 3D Stacking Neuromorphic Architecture”, 16th IEEE International Symposium on Embedded Multicore /Manycore SoCs (MCSoC-2023), Dec 18th-21th, Singapore, Singapore.

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