The University of Aizu

Mater's Thesis Research Plan

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

Ngo-Doanh NGUYEN
m5262108

2022-12-02

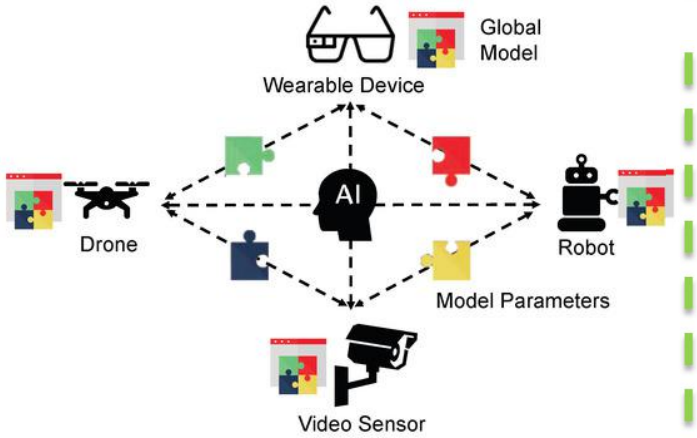
Agenda

1. Motivation & Background
2. Goals
3. Approach
4. Schedule
5. Reference

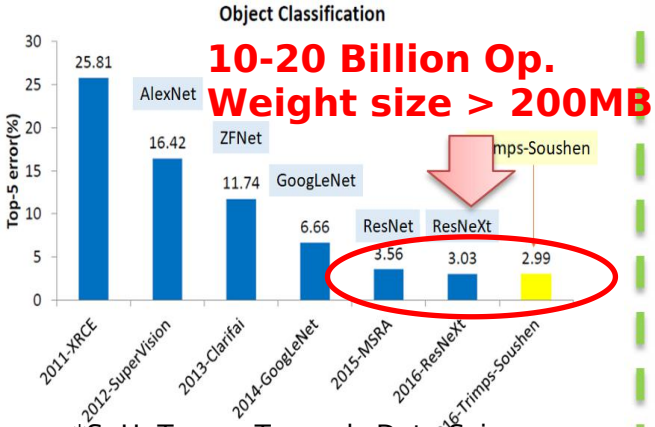
Agenda

1. Motivation & Background
2. Goals
3. Approach
4. Schedule
5. Reference

Motivation

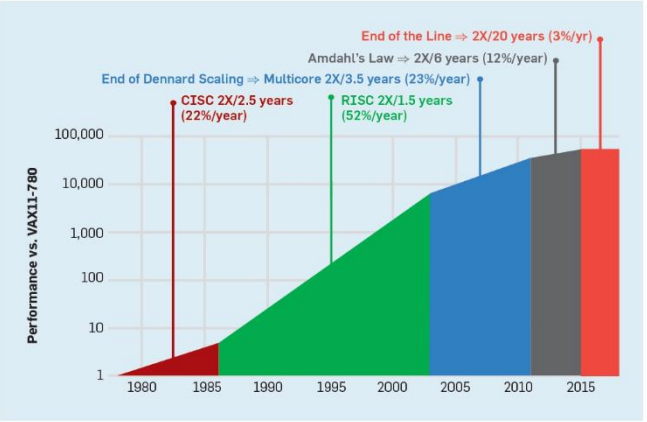


*Y. Du. Decentralized Smart IoT. Encyclopedia



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.

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

End of Moore's Law

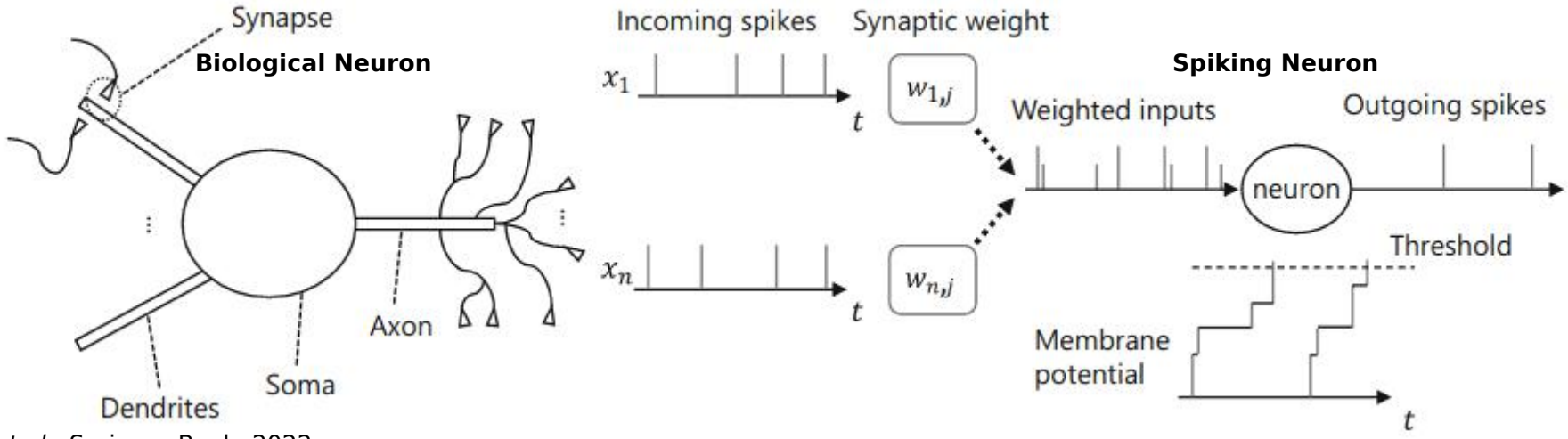
3-D Stacking Arch.

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

=> Spiking Neural Network with 3-D IC-based Stacking Memory!!!

Spiking Neural Network

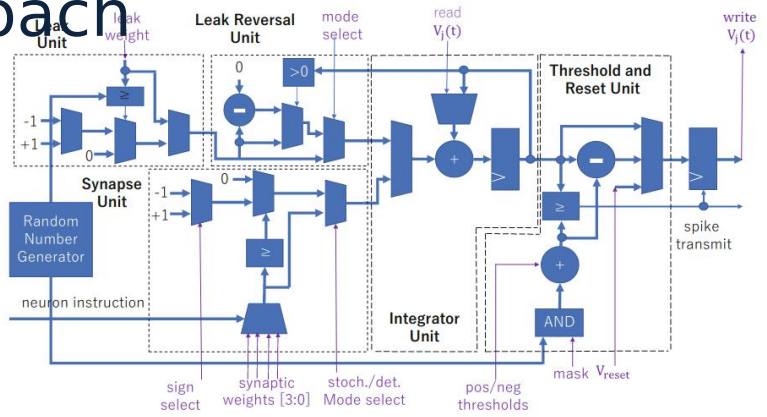
- SNNs are artifical neural network mimiced biological brian.
- There are three major parameters
  - Incoming spikes
  - Synaptic weights
  - Neuron's internal parameters (membrane potential, threshold, ...)



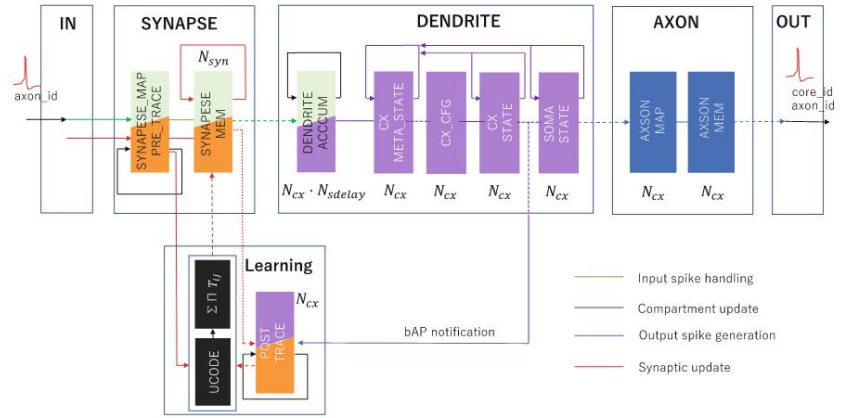
*A. Ben *et al.*, Springer Book, 2022

State of The Art (1)

- Digital Hardware Design Approach
  - IBM TrueNorth*
    - 5.4M trans. in 28-nm tech.
    - 4096 neurosynaptic cores
    - 2D Mesh 64×64 Async. NoC
  - Intel Loihi**
    - 60 mm² in 14-nm tech.
    - Scaled upto 4096 cores
    - 2D Mesh NoC



The Architecture of TrueNorth Neuron



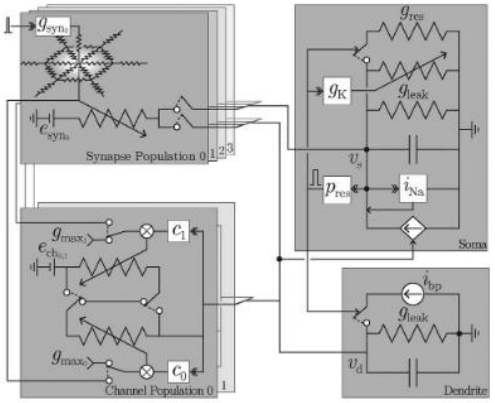
The Top-level Architecture of Loihi

NoC: Network-on-Chip
* F. Akopyan *et al.*, IEEE TCAD, 2015
** M. Davies *et al.*, IEEE Micro, 2018

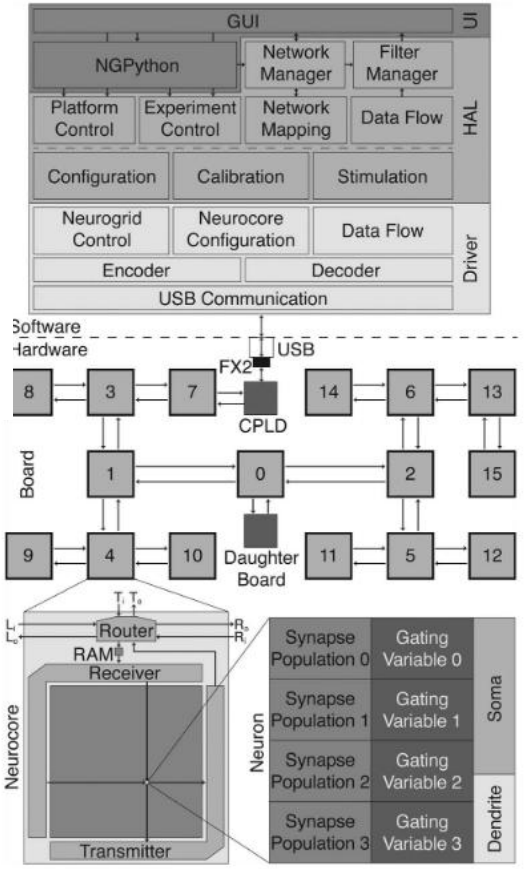
2-D NoC & Power Inefficiency for Edge devices

State of The Art (2)

- Analog and Mixed-Signal Hardware Approach
  - NeuroGrid*
    - Used a sub-threshold analog circuit for neuron
    - Digital communication using on-chip network
    - 4 ADCs to convert signals



NeuroGrid's neuron model



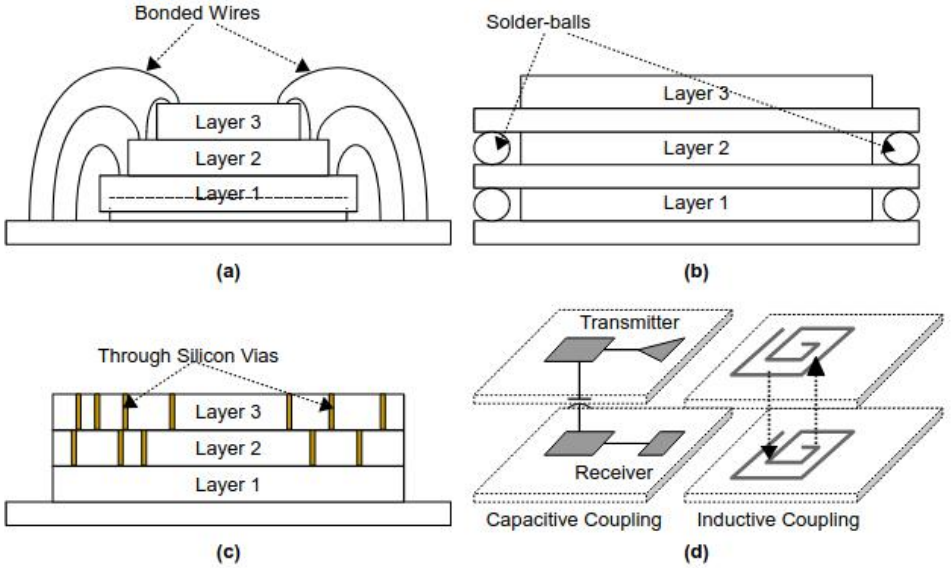
NeuroGird's software & hardwarex

2-D NoC & Lack of configurability for Edge Devices

*B. V. Benjamin *et al.*, Proceedings of the IEEE, 2014

3-D IC Opportunity

- What is 3-D IC?



(a) Wire bonding; (b) Solder balls; (c) Through Silicon Vias; (d) Wireless stacking

- Stacking multiple layers to obtain smaller footprint & shorter interconnect & power.

- Why 3-D IC?

#Bit	Kogge-Stone Adder		Log Shifter 16		Log Shifter 32	
	16-bit				32-bit	
	Delay	Power	Delay	Power	Delay	Power
2 planes	-20.2%	-8%	-13.4%	-6.5%	-28.4%	-8%
3 planes	-23.6%	-15%	-	-	-	-
4 planes	-32.7%	-22%	-	-	-	-

Performance & Power: 3D vs 2D architecture*

- #Stacking Layers ↑ ⇔ Power & Performance & Area (PPA) ↑

=> 3-D Stacking Memory for SNN is promissing

*B. Valdyanathan *et al.*, *ICVLSI Design*, 2007

Agenda

1. Motivations & Background
2. Goals
3. Approach
4. Schedule
5. Reference

Goals

- Design a hardware architecture of SNN with 3-D memory stacking
- Purpose a strategy to reduce power consumption

Agenda

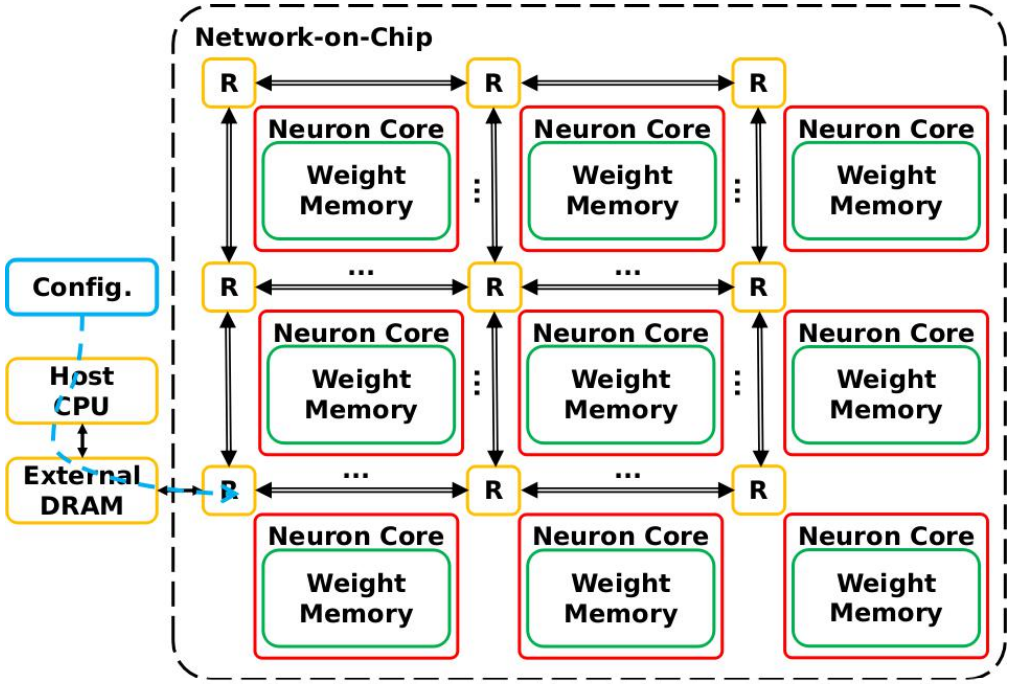
1. Motivations & Background
2. Goals
3. Approach
4. Schedule
5. Reference

Approach

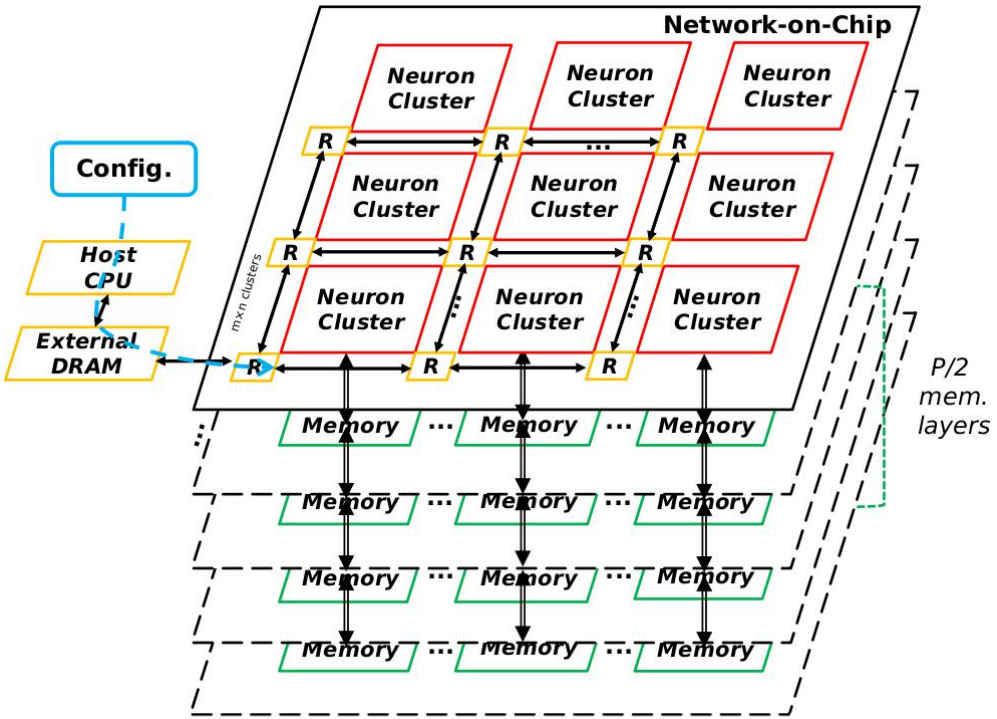
- Design a hardware architecture of SNN with 3-D memory stacking
  - Use **Through-Silicon-Vias (TSV)** for 3-D SNN architecture with configurable feature.
- Purpose a strategy to reduce power consumption
  - Use a strategy to **Turn-on/off inefficiency memory layers.**

Hardware Architecture (1)

- Use **Through-Silicon-Vias (TSV)** for 3-D SNN arch.



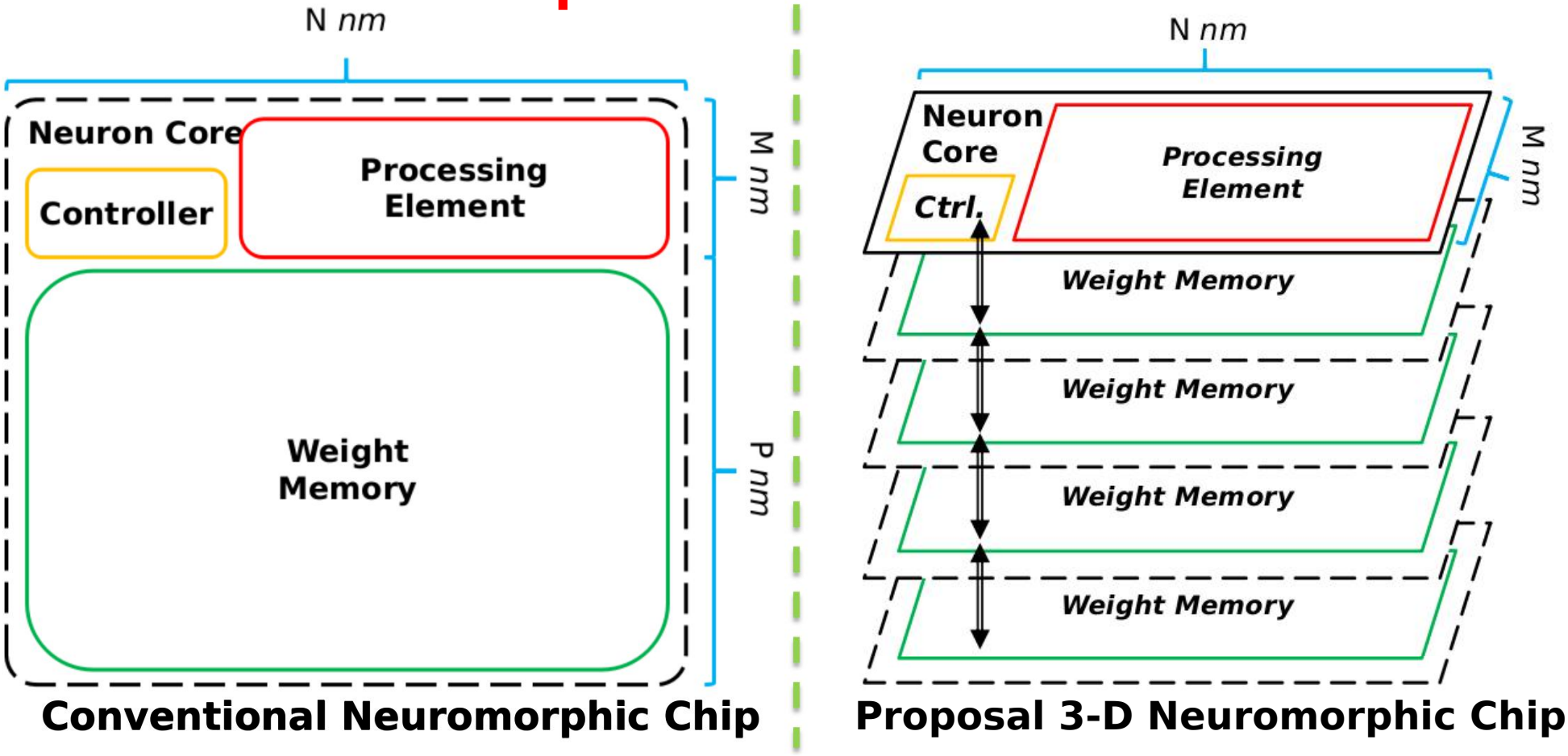
Conventional Neuromorphic Chip



Proposal 3-D Neuromorphic Chip

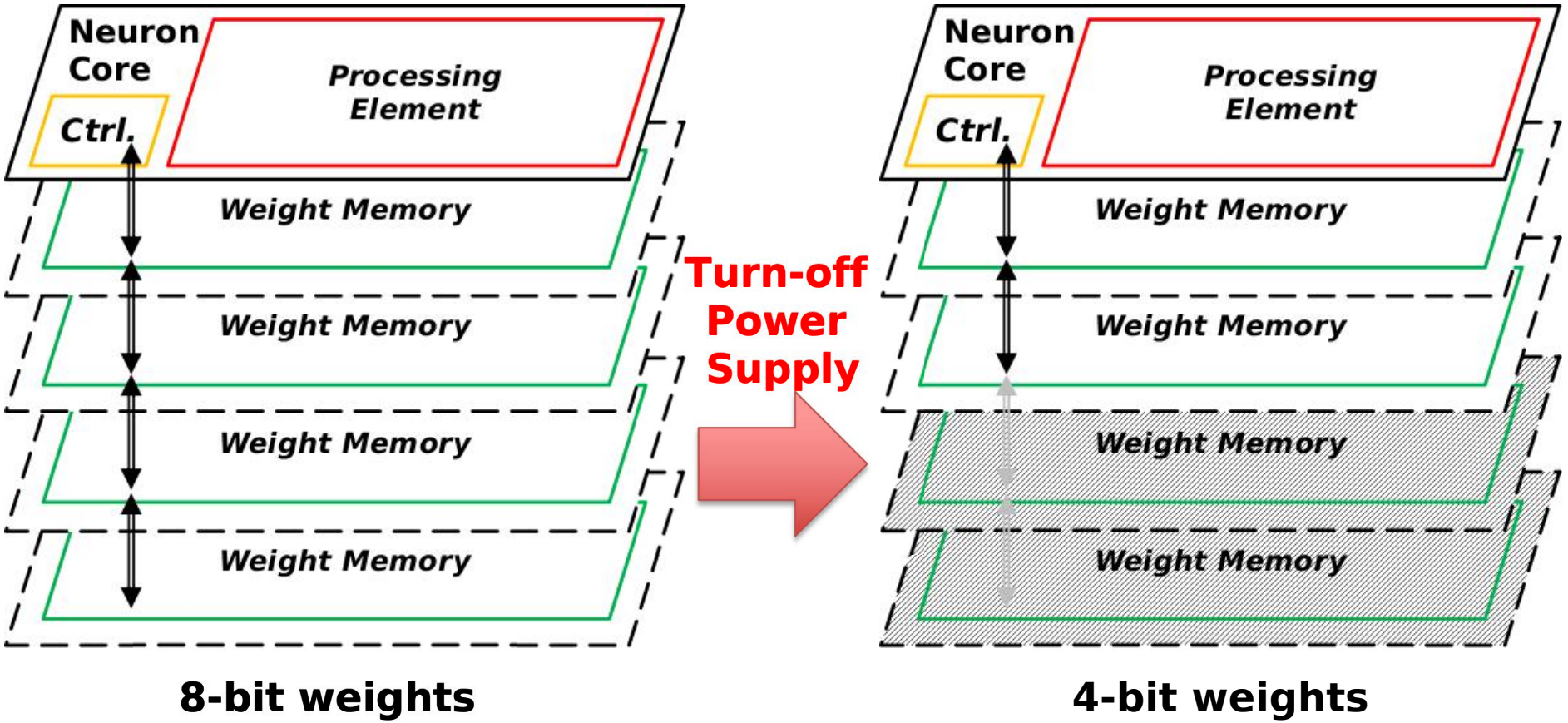
Hardware Architecture (2)

- **Hardware footprint** can be **reduced** with 3-D stacking



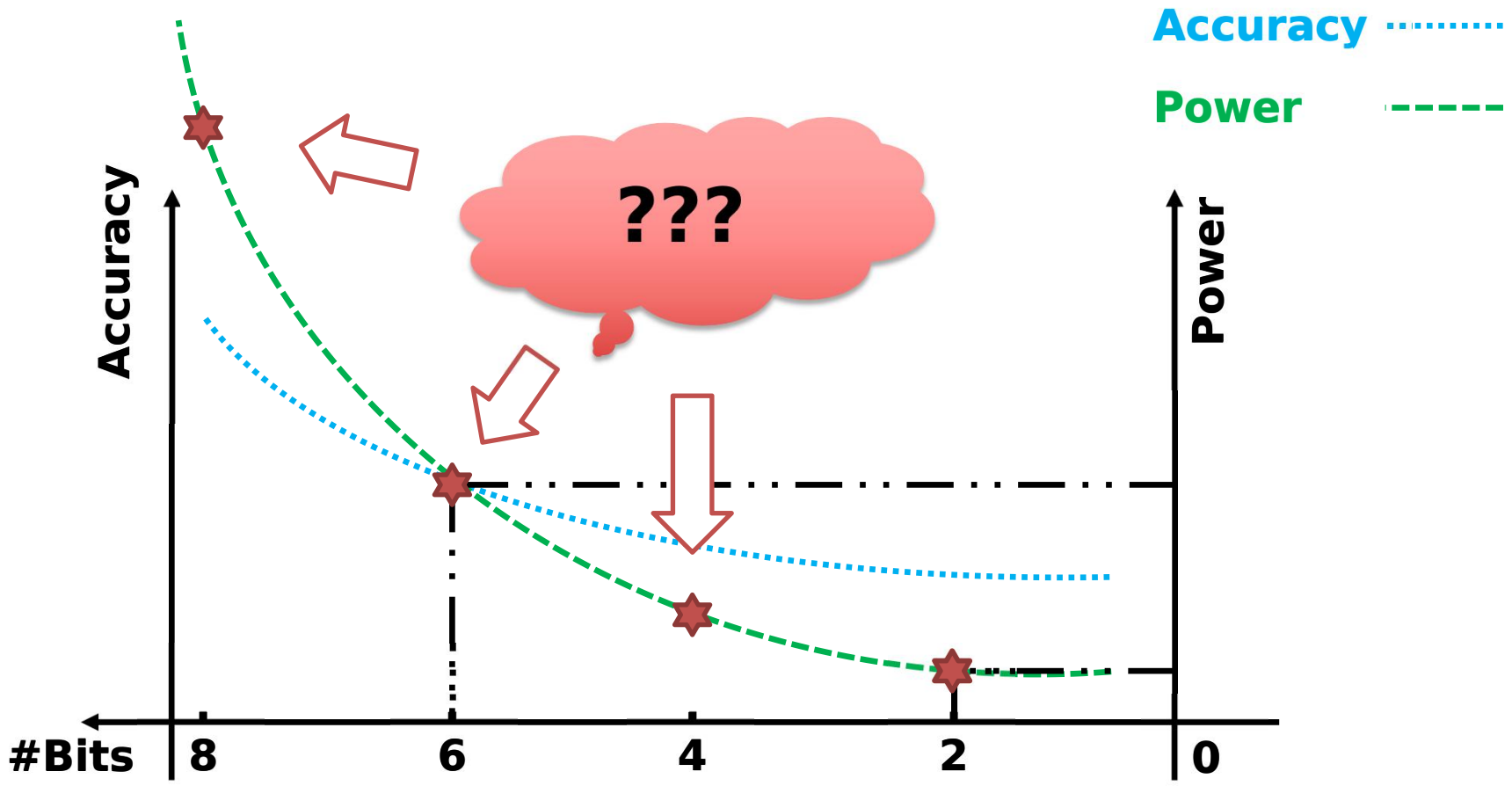
Power Consumption Reduction Strategy (1)

– Use a strategy to **Turn-on/off inefficiency memory layers.**



Power Consumption Reduction Strategy (2)

- Find **the optimal point** between Accuracy & Power



Agenda

1. Motivation
2. Goal
3. Approach
4. Schedule
5. Reference

Road Map

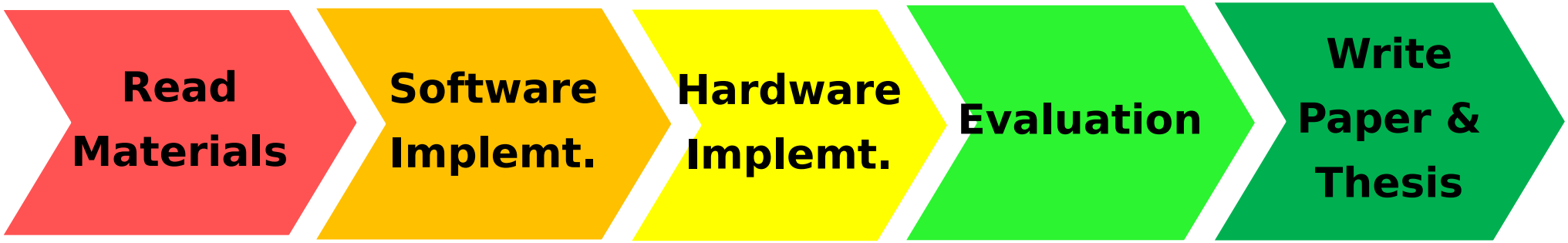
- Master's Course Schedule

1/23 - 5/23

- Build SW model as a golden reference for HW

12/23 - 3/24

- Evaluation power consumption & accuracy



10/22 - 1/23

- Get familiar with SNN

5/23 - 12/23

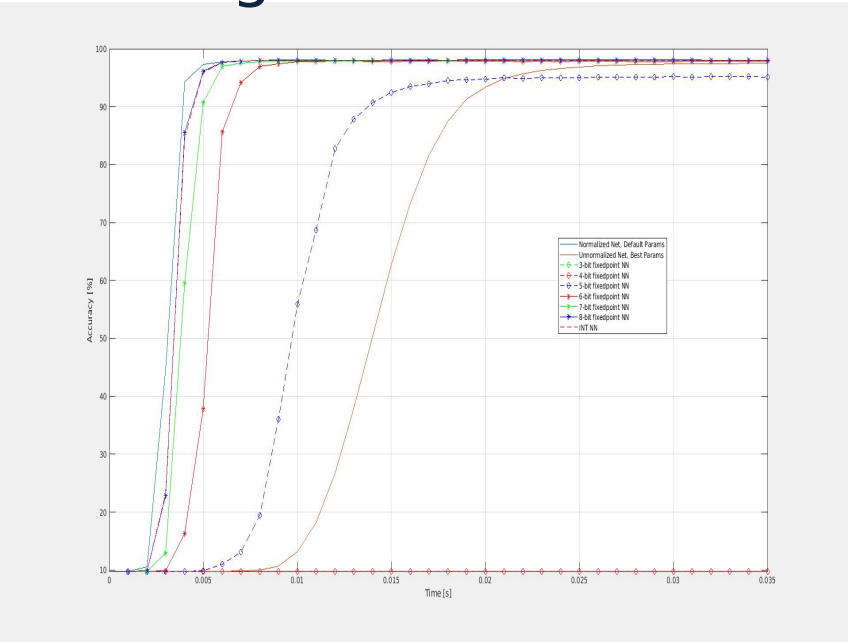
- Build HW model
- Optimize HW model

3/24 - 9/24

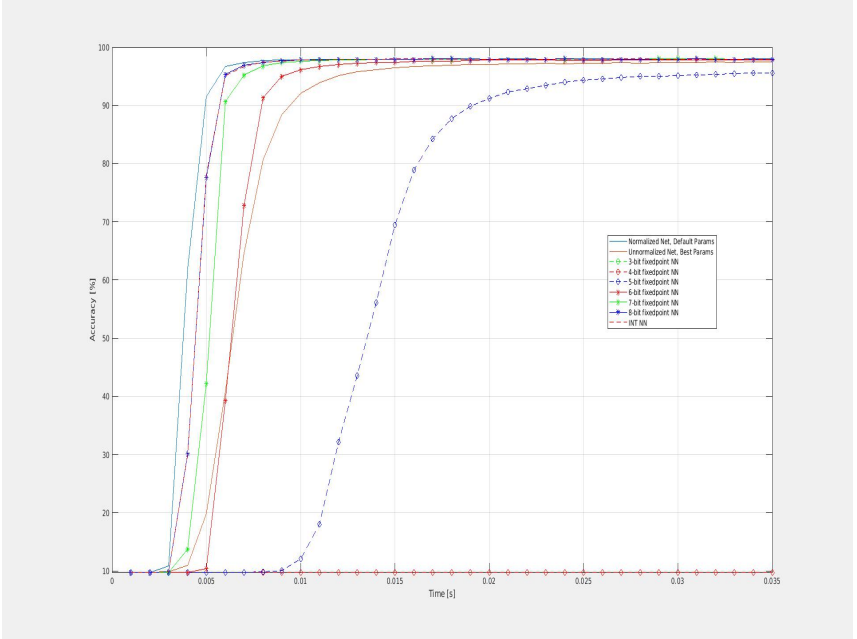
- Submit papers
- Complete thesis

Progress (1)

- Understand basic idea of SNN algorithm
- Run sample Software model with different configurations



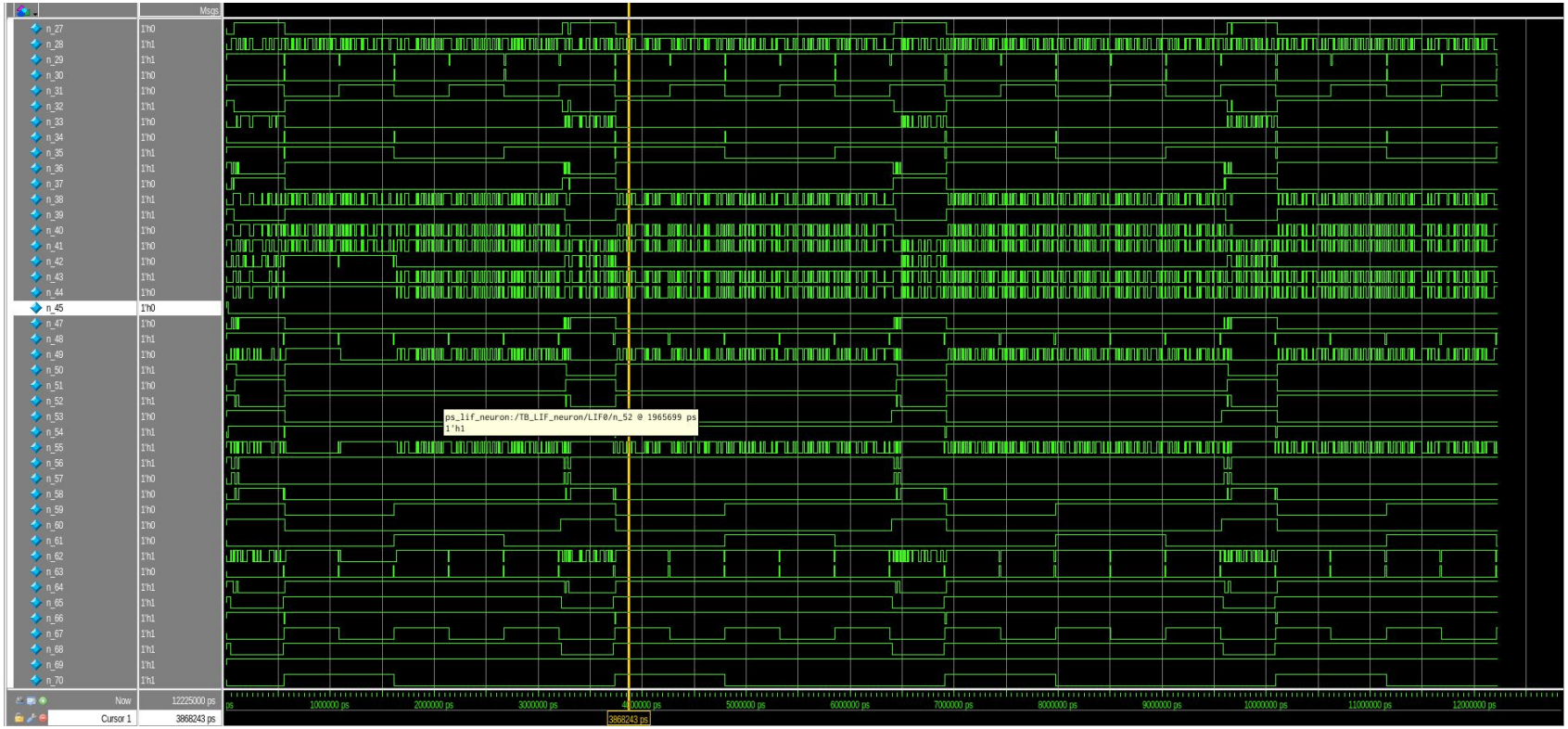
Layer Config. = [784 1200 512 84 10]
Learning Rate = 0.5
#Epoch = 15



Layer Config. = [784 512 84 10]
Learning Rate = 1
#Epoch = 15

Progress (2)

- Simulate a simple processing element of SNN (LIF Core)



RTL simulation of LIF Core

Agenda

1. Motivation
2. Goal
3. Approach
4. Schedule
5. Reference

Reference

1. Du, Y. Decentralized Smart IoT. Encyclopedia. Available online: https://encyclopedia.pub/entry/8977 (accessed on 22 November 2022).
2. S. H. Tsang, https://towardsdatascience.com/review-refinenet-multi-path-refinement-network-semantic-segmentation-5763d9da47c1
3. John L. Hennessy, David A. Patterson, A New Golden Age for Computer Architecture, Communications of the ACM, February 2019, Vol. 62 No. 2, Pages 48-60, 10.1145/3282307
4. A. Ben, K. N, Dang, Neuromorphic Computing Principles and Organization, Springer, 2022
5. F. Akopyan et al., "TrueNorth: Design and Tool Flow of a 65 mW 1 Million Neuron Programmable Neurosynaptic Chip," in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 34, no. 10, pp. 1537-1557, Oct. 2015, doi: 10.1109/TCAD.2015.2474396.
6. M. Davies et al., "Loihi: A Neuromorphic Manycore Processor with On-Chip Learning," in IEEE Micro, vol. 38, no. 1, pp. 82-99, January/February 2018, doi: 10.1109/MM.2018.112130359.
7. B. V. Benjamin et al., "Neurogrid: A Mixed-Analog-Digital Multichip System for Large-Scale Neural Simulations," in Proceedings of the IEEE, vol. 102, no. 5, pp. 699-716, May 2014, doi: 10.1109/JPROC.2014.2313565.
8. B. Vaidyanathan, W.-L. Hung, F. Wang, Y. Xie, V. Narayanan, and M. J. Irwin, “Architecting microprocessor components in 3d design space,” in 20th International Conference on VLSI Design, 2007. Held Jointly with 6th International Conference on Embedded Systems., pp. 103–108, IEEE, 2007.

The University of Aizu

Thank you

for your attention.