Master's Thesis Research Plan

Efficient Hardware Implementation using Izhikevich Neuron and 3-D based Stacking Memory

SUBBAIAH RAVI Hariprakash

m5282026

29/11/2024

Agenda

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

Agenda

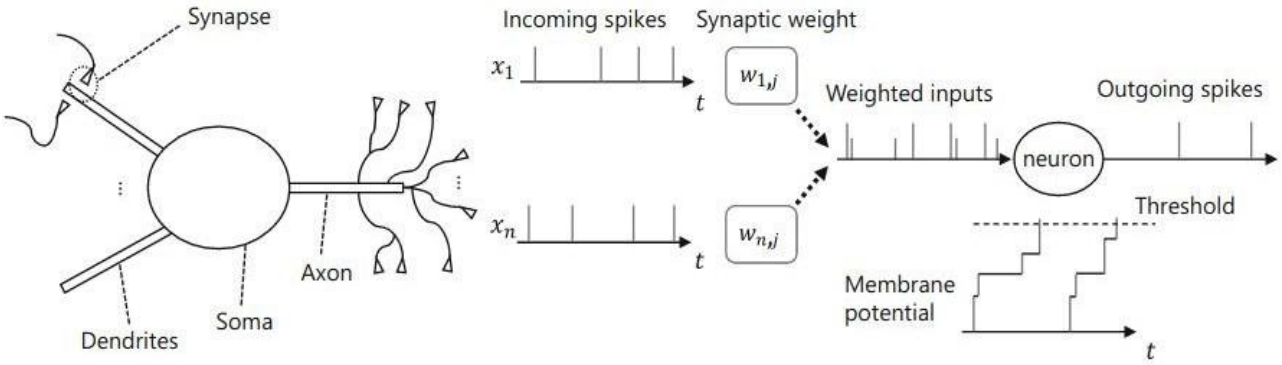
1. Motivation & Background
2. Goals
3. Approach
4. Schedule
5. Reference

Background

- Increasing computations increases power consumption
- Spiking Neural Networks are energy efficient
- 3-D stacking memory and approximate computation are promising solutions

Spiking Neural Network

- Energy-efficient artificial neural networks that process information using discrete events or spikes, mimicking biological neurons

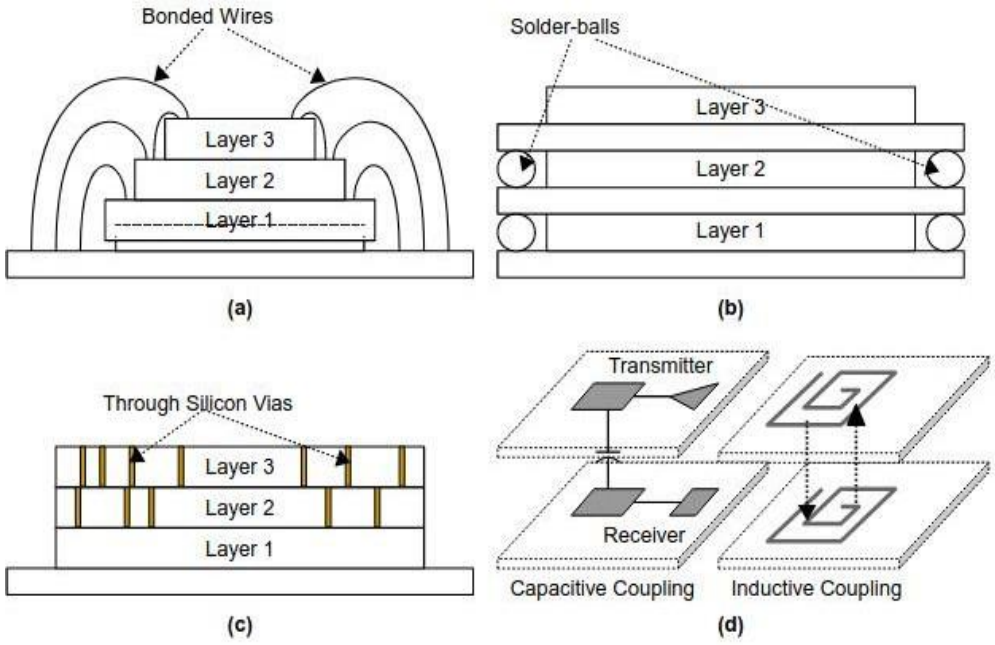


- Major parameters
  - Incoming spikes
  - Synaptic weights
  - Neuron's internal parameters

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

3-D IC Stacking

- Stacking multiple layers vertically
- Smaller footprint
- Shorter interconnections
- Low power



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

Approximate Computing

- Replace accurate adders and multipliers with approximate adders and multipliers.
- Computation error arises from approximation
- SNN is resistant to such noise.

Circuit name	MAE	WCE	MRE	EP	power	area
mul8u_1JJQ [1]	0.00 %	0.00 %	0.00 %	0.00 %	0.391	709.6
mul8u_2V0 [1]	0.0015 %	0.0046 %	0.052 %	64.06 %	0.386	676.3
mul8u_LK8 [1]	0.0046 %	0.017 %	0.18 %	75.00 %	0.370	637.8
mul8u_R92 [1]	0.017 %	0.061 %	0.59 %	87.54 %	0.345	604.5
mul8u_0AB [1]	0.057 %	0.18 %	2.56 %	97.72 %	0.302	542.5

8-bit Unsigned Multiplier comparison
Reported error parameters: MAE - Mean Absolute Error (Mean Error Magnitude), WCE - Worst-Case Absolute Error (Error Magnitude / Error Significance), MSE - Mean Squared Error, MRE - Mean Relative Error (Mean Relative Error Distance), WCRE - Worst-Case Relative Error, EP - Error Probability (Error Rate) | Reported design parameters: power - power consumption in mW, area - area on the chip in um2

https://ehw.fit.vutbr.cz/evoapproxlib/

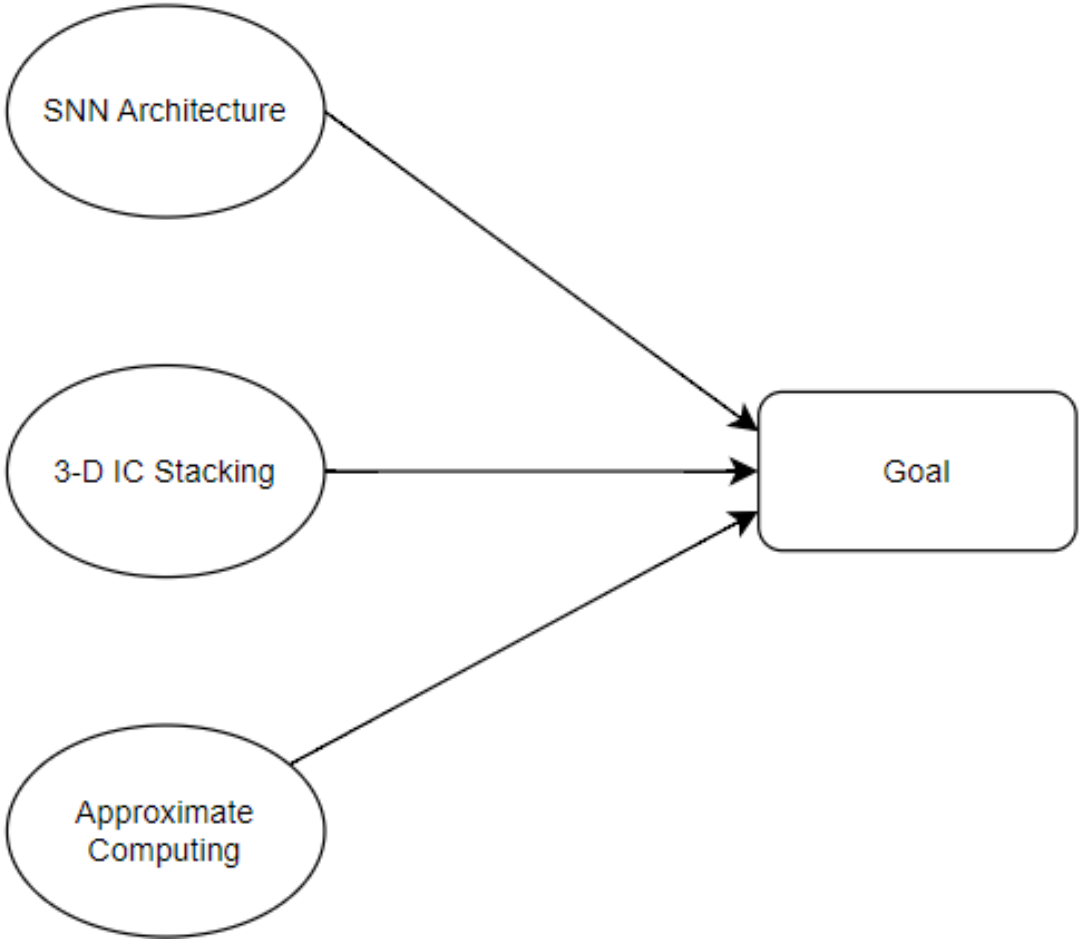
Motivation

- A more energy efficient SNN using:
  - Approximate computing: Lower power, smaller area
  - 3-D stacking IC: Reduce footprint, lower power
- How to combine both strategies?

Agenda

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

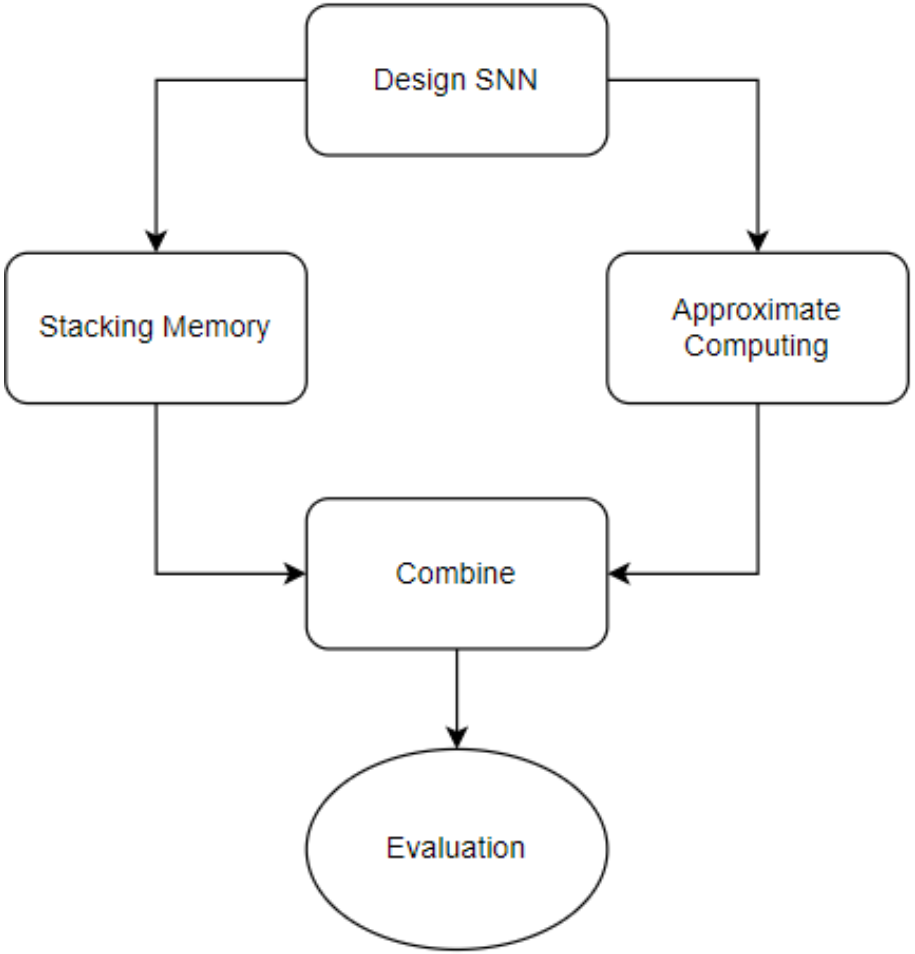
Goals


SNN Architecture
3-D IC Stacking
Approximate Computing
Goal

Agenda

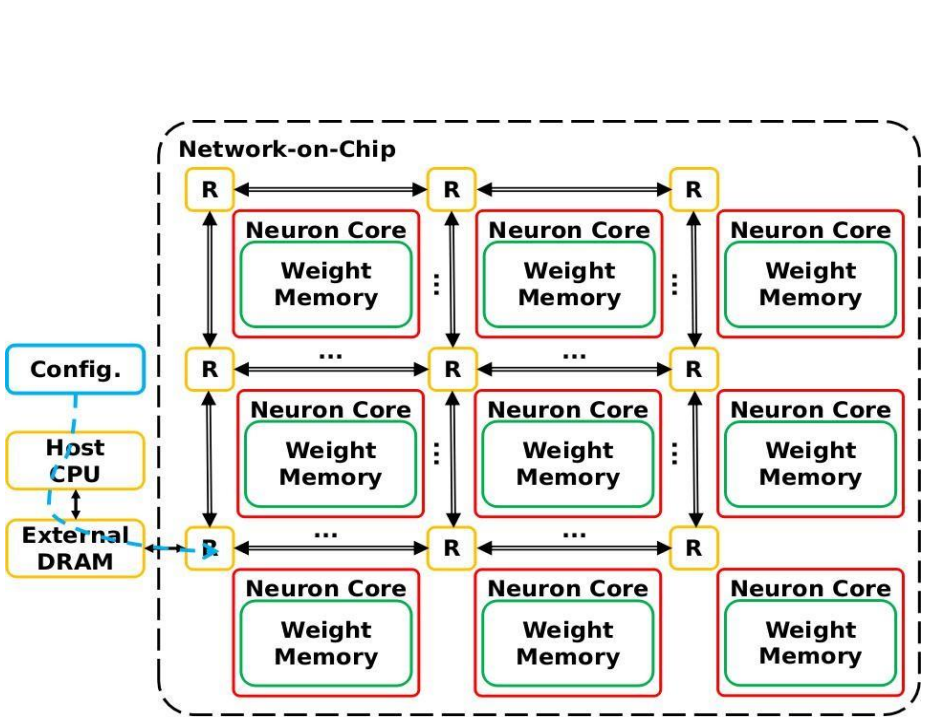
1. Motivations & Background
2. Goals
3. Approach
4. Schedule
5. Reference

Approach

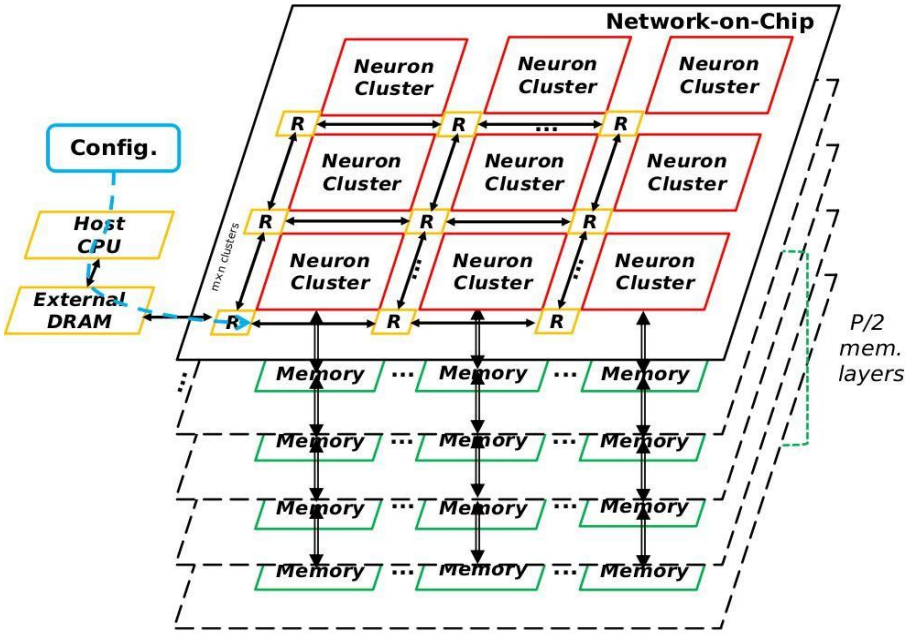

Design SNN
Stacking Memory
Approximate Computing
Combine
Evaluation

TSV Stacking Structure

- Use Through-Silicon-Vias (TSV) for 3-D SNN architecture [11]

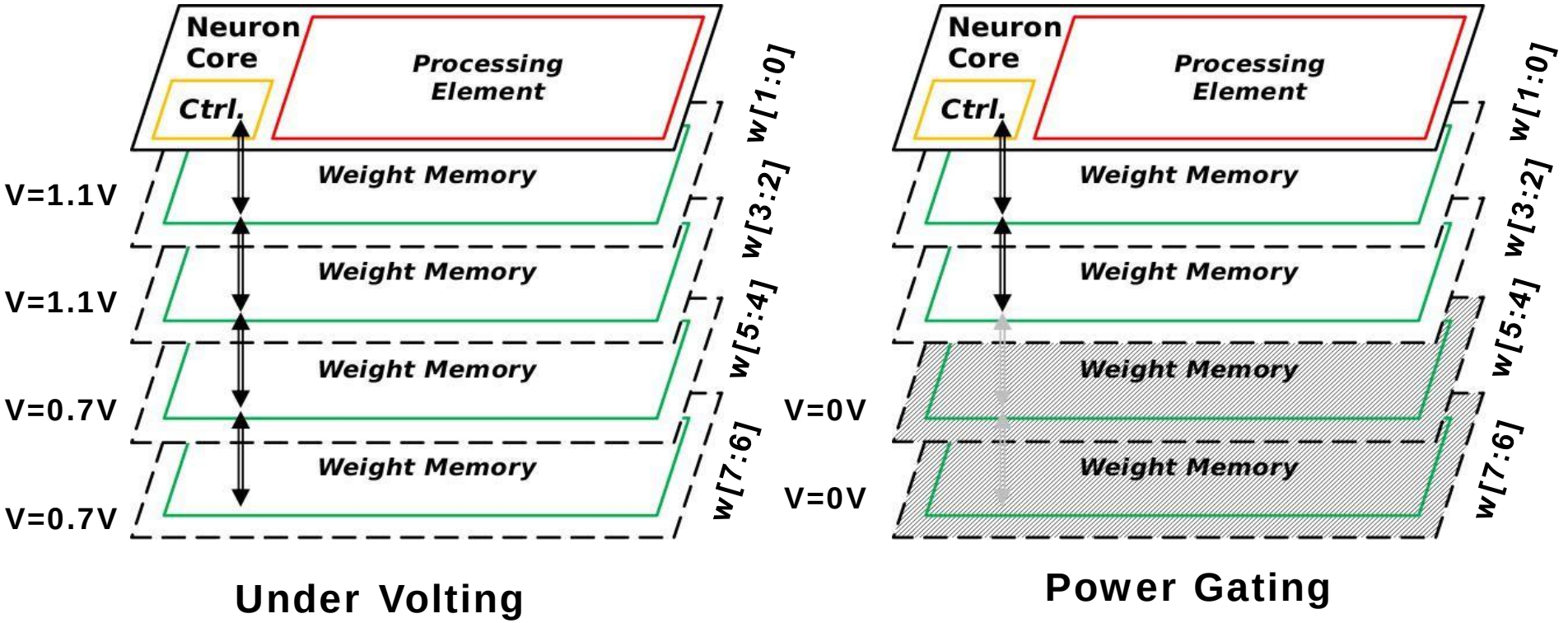


Conventional Neuromorphic Chip



Our 3-D Neuromorphic Chip

Low Power Strategies



- Split weight w[7:0] into 4 subsets and put into different layers
- Undervolt and powergate LSBs

Implementing Approximation (2)

$$\frac{dv}{dt} = 0.04v^2 + 5v + 140 - u + I$$
$$\frac{du}{dt} = a(bv - u)$$
$$\text{if } v \geq 30mV, \text{ then } \begin{cases} v \to c \\ u \to u + d \end{cases}$$

Original Izhikevich Equation

$$\frac{dv}{dt} = 2^{-2}v^2 + 2^2v + v + 14 - u + I$$
$$\frac{du}{dt} = 2^{-6}(2^{-2}v - u)$$
$$\text{if } v \geq 3mV, \text{ then } \begin{cases} v \to c \\ u \to u + d \end{cases}$$

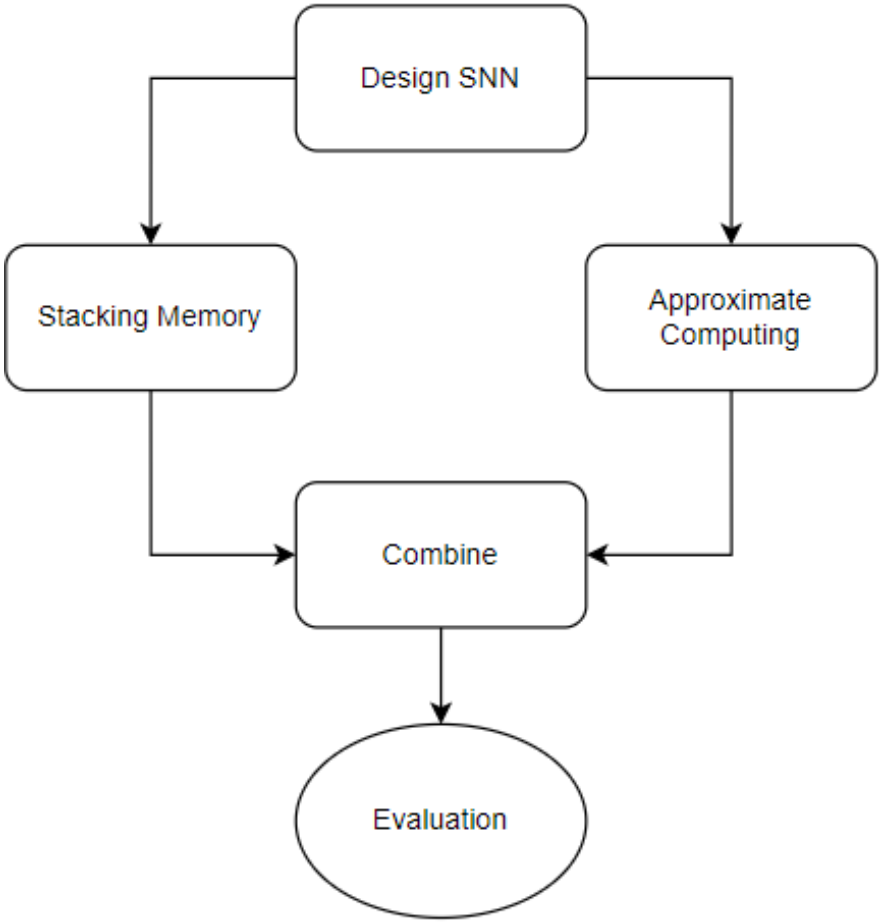
Modified HOMIN Equation

- [1] R. Kobayashi and K. N. Dang, "An Efficient Hardware Implementation of Spiking Neural Network Using Approximate Izhikevich Neuron," 2024 9th International Conference on Integrated Circuits, Design, and Verification (ICDV), Hanoi, Vietnam, 2024, pp. 13-18, doi: 10.1109/ICDV61346.2024.10616602. keywords: {Accuracy;Power demand;Costs;Computational modeling;Neurons;Spiking neural networks;Hardware;Approximate computing;Spiking neural network;Izhikevich neuron model;FPGA;Classification},

Approach

Evaluation:

- Power consumption
- Area cost
- Accuracy
- Tradeoff b/w power/area/accuracy



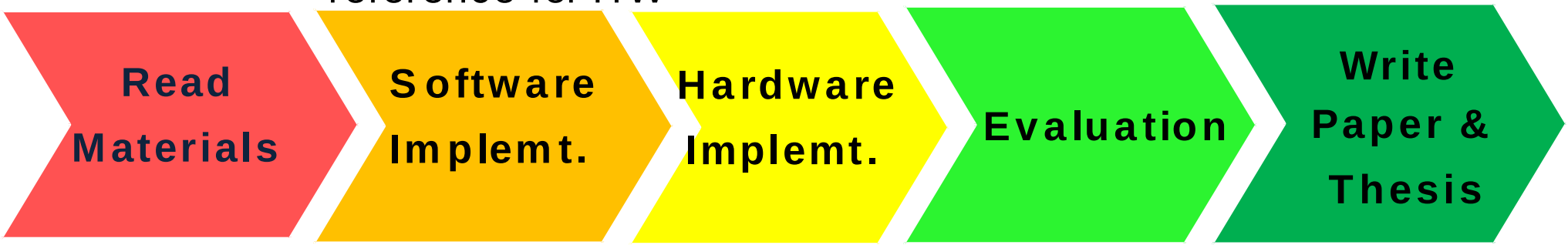
Agenda

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

Schedule

1/25 - 5/25
- Build SW model as a golden reference for HW

12/25 - 3/26
- Evaluation power consumption & accuracy



12/24 - 1/25
- Get familiar with SNN

5/25 - 12/25
- Build HW model
- Approximate Computing
- 3-D IC Stacking

3/26 - 9/26
- Submit papers
- Complete thesis

Agenda

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

Reference (1)

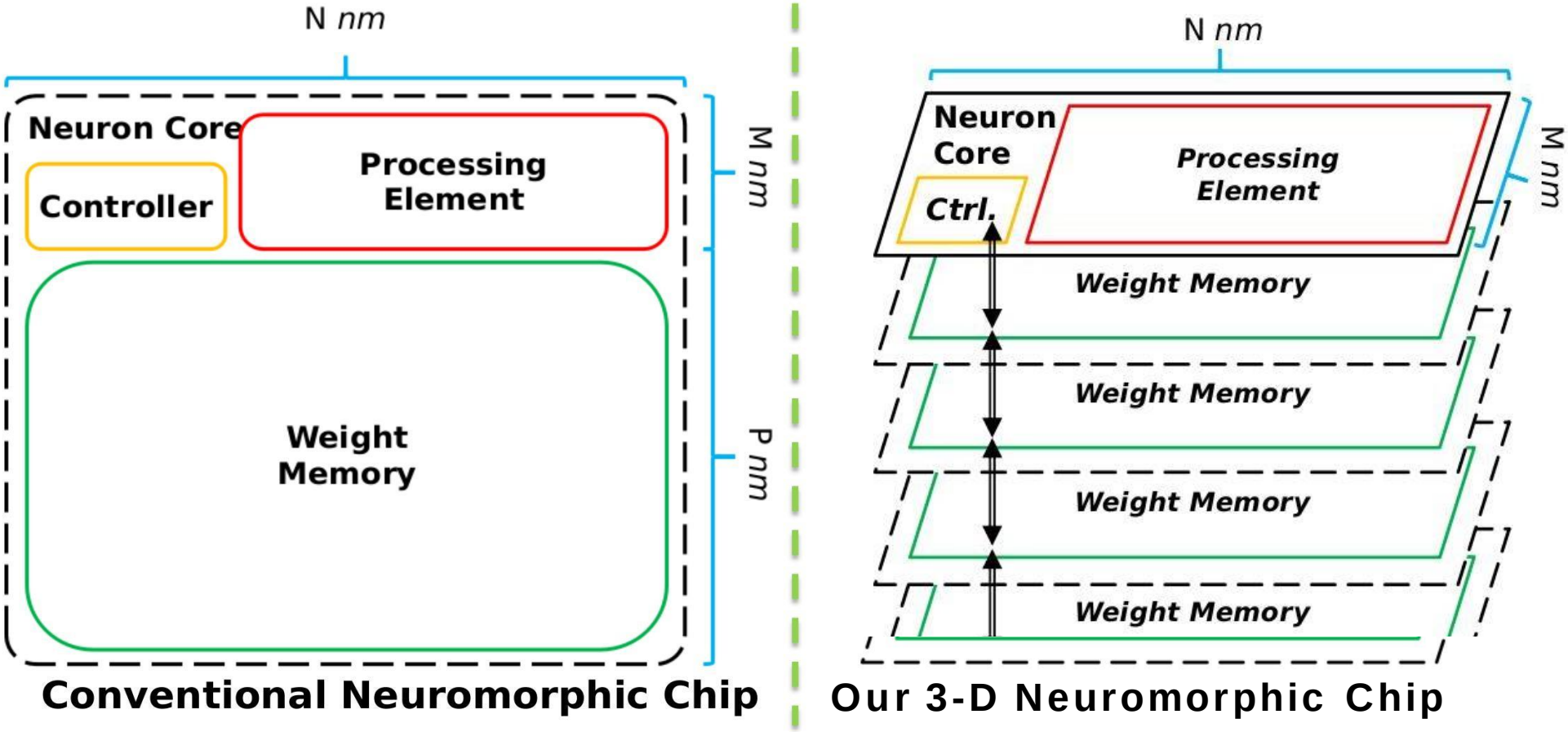
- [1] R. Kobayashi and K. N. Dang, "An Efficient Hardware Implementation of Spiking Neural Network Using Approximate Izhikevich Neuron," 2024 9th International Conference on Integrated Circuits, Design, and Verification (ICDV), Hanoi, Vietnam, 2024, pp. 13-18, doi: 10.1109/ICDV61346.2024.10616602.
- [2] Ryoji Kobayashi, Ngo-Doanh Nguyen, Nguyen Anh Vu Doan and Khanh N. Dang, "Energy-Efficient Spiking Neural Networks Using Approximate Neuron Circuits and 3D Stacking Memory‘‘, 2024 IEEE 17th International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSoC), Dec. 16-19, 2024.
- [3] Du, Y. Decentralized Smart IoT. Encyclopedia. Available online: https://encyclopedia.pub/entry/8977 (accessed on 22 November 2022).
- [4] S. H. Tsang, https://towardsdatascience.com/review-refinenet-multi-path-refinement- network-semantic-segmentation-5763d9da47c1
- [5] 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

Reference (2)

- [6] A. Ben, K. N, Dang, Neuromorphic Computing Principles and Organization, Springer, 2022
- [7] 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.
- [8] 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.
- [9] 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.
- [10] 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.
- [11] 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, vol. 31, no. 12, pp. 2016-2029, Dec. 2023, doi: 10.1109/TVLSI.2023.3318231

Thank You

Hardware Footprint



Poisson Spikes

- MNIST database is 8-bit so Poisson spike trains are generated for neuron input.