University of Aizu

会津大学
THE UNIVERSITY OF AIZU

Master’s Research Plan Presentation

# Dynamic Quantization & Pruning in Spiking Neuron Networks


Yassine Khedher

m5281019


July 24th 2024

# Content

# Content

1. Motivation & Background
2. Research Goal
3. Approach/Method
4. Schedule

# Need for new Architectures

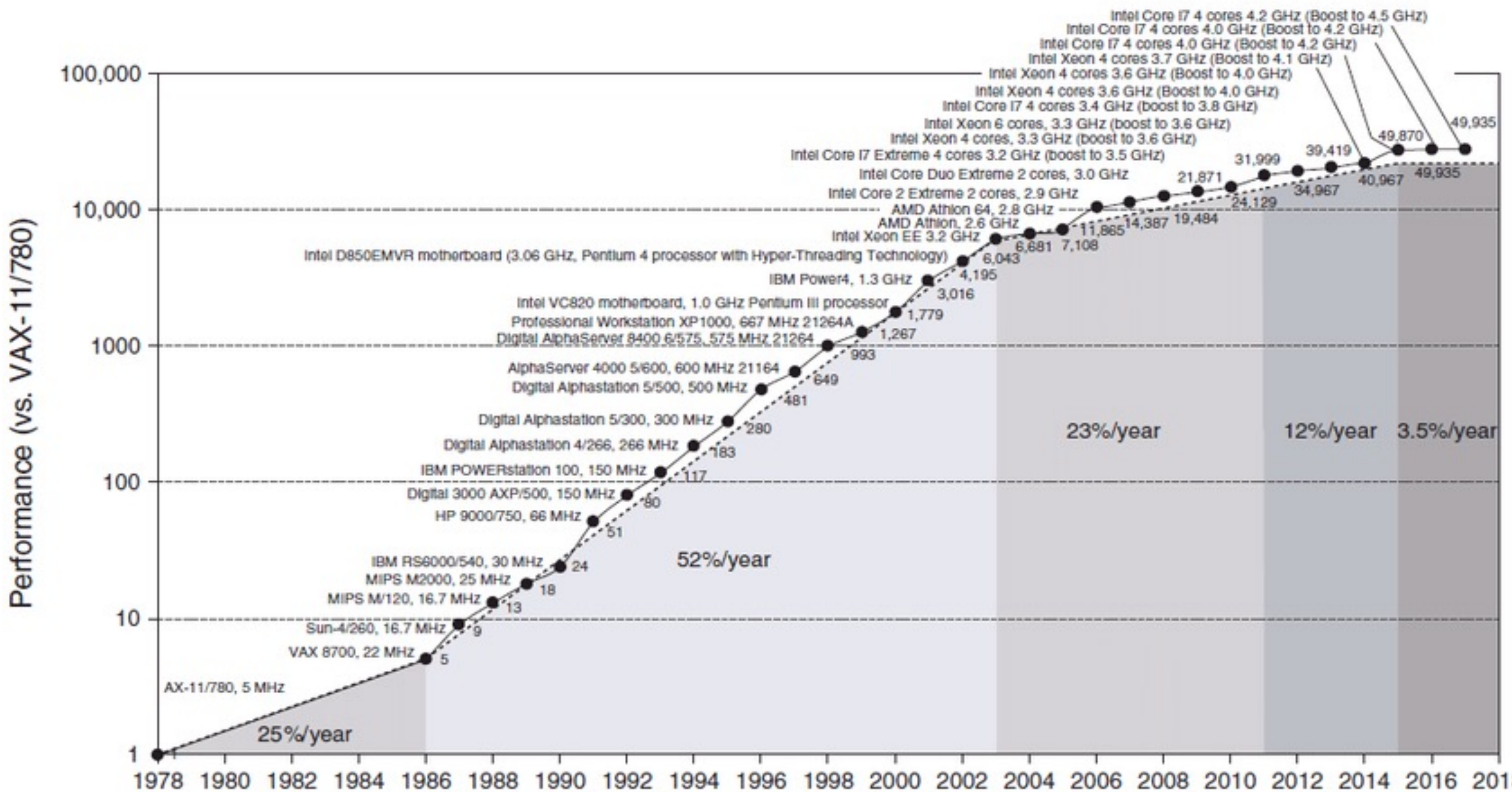



Figure: Growth in processor performance over 40 years (Moore's Law) [1]

## → Neuromorphic Computing

- Inspired by the human brain's structure and function
- Offers new ways to achieve high efficiency and performance

# Neuromorphic Computing

## Why Neuromorphic Computing?

- Suitable for applications requiring low power and real-time processing
- Aims to create hardware and algorithms that mimic neural processes

## Advantages

- Highly efficient in terms of power consumption
- Event-driven computing
- Potential for robust and scalable AI applications

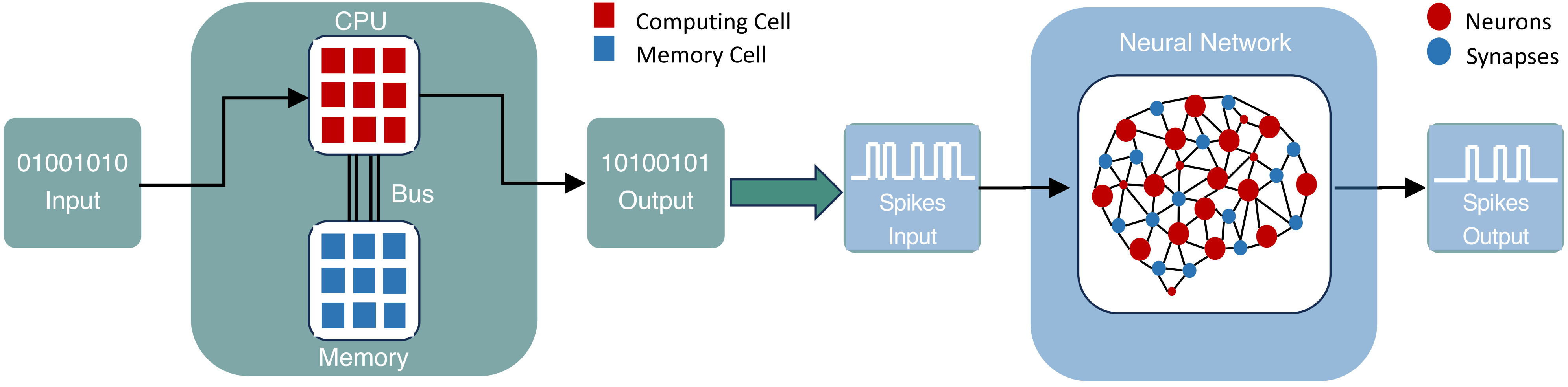



Figure: Differences between Conventional and Neuromorphic Computing

# Spiking Neuron Networks (SNNs)

**What are SNNs?**

- More biologically realistic compared to traditional neural networks
- Use discrete spikes to represent and process information

**Advantages:**

- Efficient in terms of power and data processing
- Capable of learning temporal patterns and sequences
- Better suited for real-time and low-power applications

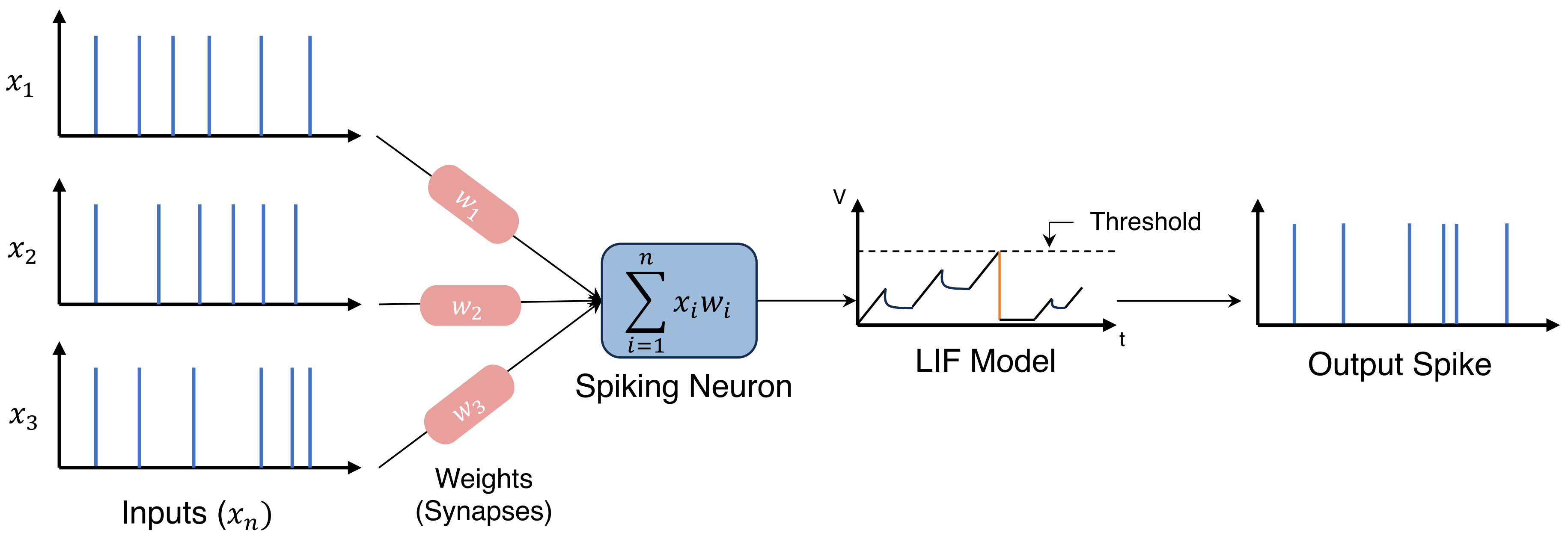



Figure: Flow of data in a Spiking Neuron Network

# Related Works

## Model compression techniques

Table: State-of-the-art of model compression techniques in SNNs

| Ref. | Technique | Accuracy | Energy Consumption | Limitations | Dataset |
|---|---|---|---|---|---|
| [2] | Knowledge Distillation | 74.42 % | 3,3x improvement | Relies on a well-trained teacher network | MNIST |
| [3] | Static Pruning | 97.57 % | 3.1x improvement | Iso-accuracy maintained only up to 80% sparsity | DVS Gestures |
| [3] | Static Quantization | 87.85% | 2.4x energy improvement | Acc. loss at low bit-widths | DVS Gestures |
| [3] | Joint Quantization & Pruning | - | 10x improvement in energy-delay product (EDP) | May result in accuracy loss | DVS Gestures |
| [4] | STDP Pruning & Weight Quantization | MNIST: 90.1%<br>Caltech-101: 91.6% | MNIST: 3.1x improvement<br>Caltech-101: 2.2x improvement | Accuracy sensitive to pruning threshold and number of quantization | MNIST<br>Caltech-101 subset |

## Existing Gaps and Limitations:

- Limited integration of dynamic quantization & pruning with SNNs.
- Lack of comprehensive frameworks addressing both efficiency and resource constraints.

# Content

# Challenges in SNNs

**Reducing the energy consumption:**

- Increasing the power efficiency can reduce the accuracy

→ Tradeoff between accuracy and energy consumption
→ How to reduce energy consumption without significantly affecting the accuracy?

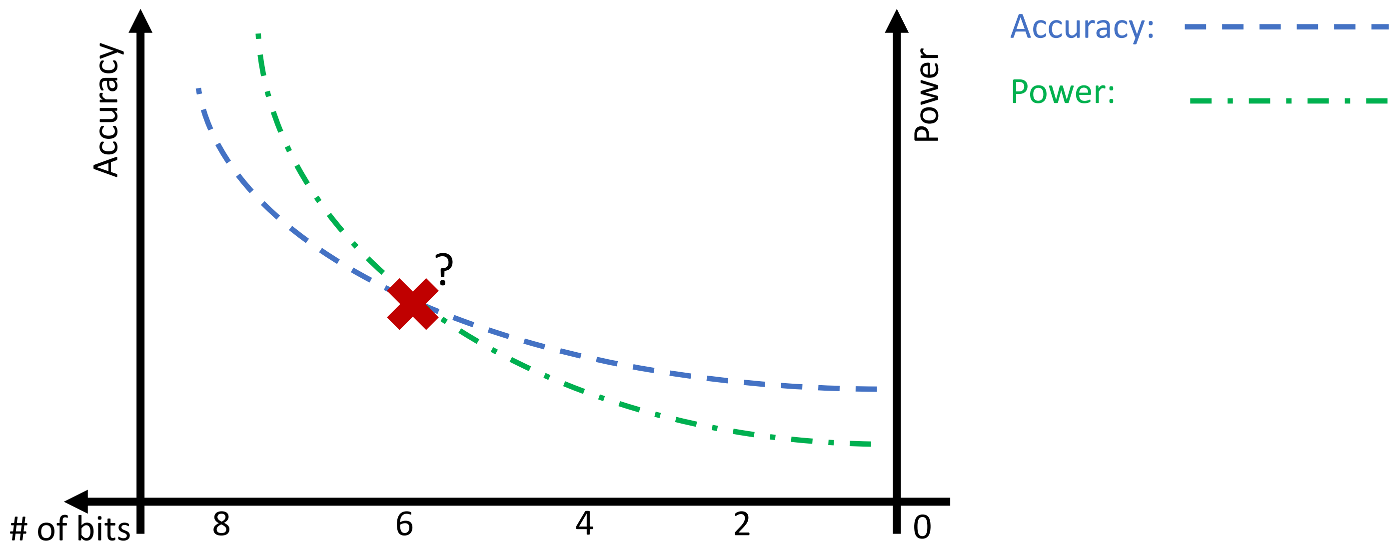



Figure: Relation between Accuracy, Power consumption and Number of bits in SNN

# Goals

**Objective:**

- Develop and evaluate methods for Dynamic Quantization and Pruning in Spiking Neuron Networks (SNNs).

**Goals:**

- Implement Dynamic Quantization & Pruning in order to:
  - → Optimize Energy Consumption
  - → Improve Computational Efficiency
  - → Maintain or Enhance Accuracy
  - → Develop a Comprehensive Evaluation Framework
  - → Contribute to the Field of Neuromorphic Computing

# Content

# Dynamic Quantization

**Definition:**

- Dynamic quantization involves adjusting the precision of weights and activations in neural networks during runtime to reduce computational load and energy consumption.

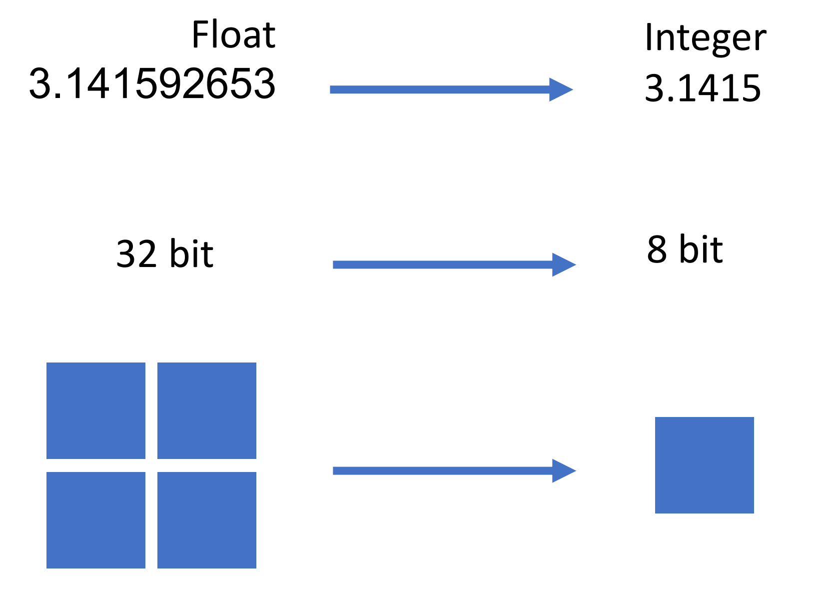



Figure: Examples of quantization process

# Dynamic Pruning

**Definition:**

- Involves selectively deactivating neurons and synapses during runtime based on their activity levels, thus reducing the network's complexity and energy consumption.

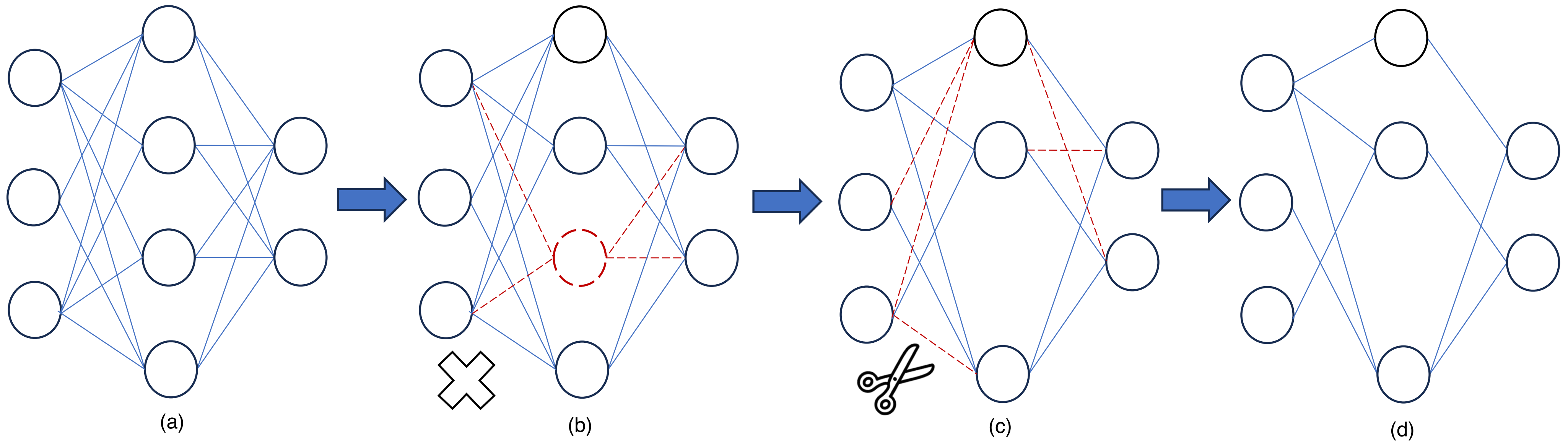


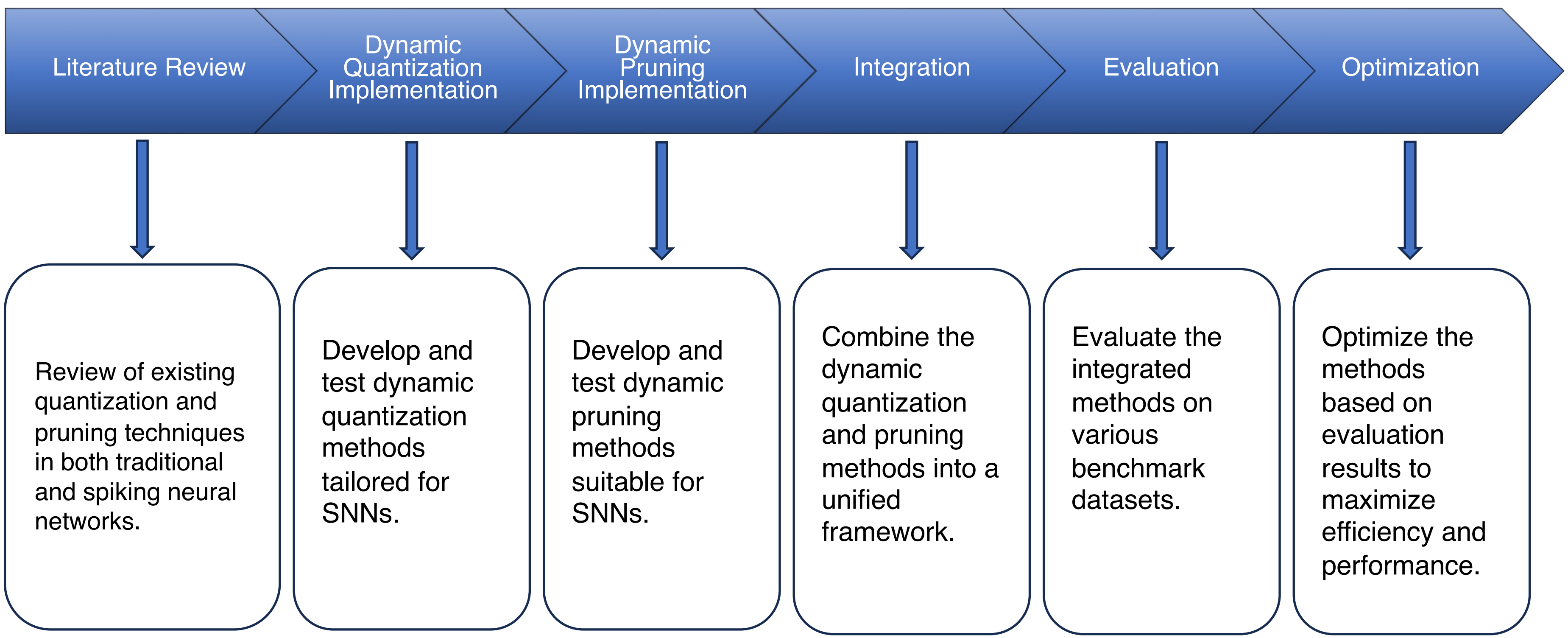

Figure: Pruning Examples in Neural Networks: Neurons Pruning (b) & Synapses Pruning (d)

# Approach

**Overview:**

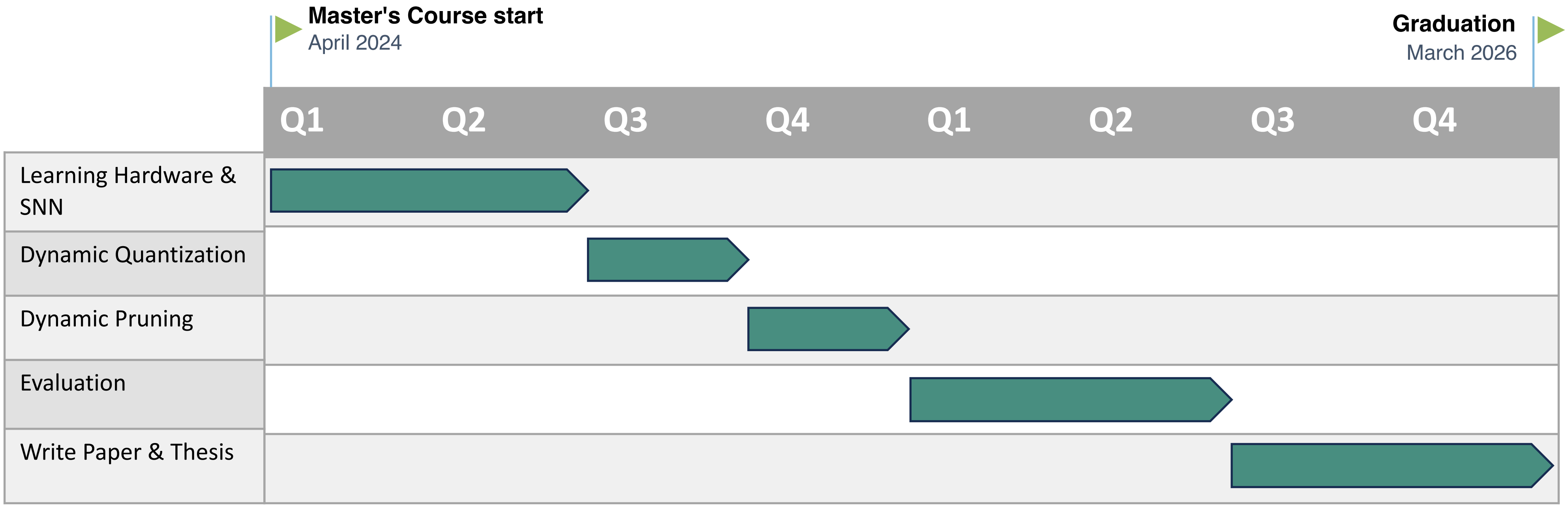
- Developing and integrating dynamic quantization and pruning techniques into Spiking Neuron Networks (SNNs) to optimize their performance.

**Steps:**



# Content

# Schedule


Master's Course start
April 2024
Graduation
March 2026
Q1
Q2
Q3
Q4
Q1
Q2
Q3
Q4
Learning Hardware & SNN
Dynamic Quantization
Dynamic Pruning
Evaluation
Write Paper & Thesis

# References


- [1] HENNESSY, D. A. P. J. L. Computer Architecture, Sixth Edition: A Quantitative Approach. 6. ed. [S.l.]: Morgan Kaufmann, 2017. (The Morgan Kaufmann Series in Computer Architecture and Design).
- [2] S. Takuya, R. Zhang and Y. Nakashima, "Training Low-Latency Spiking Neural Network through Knowledge Distillation," 2021 IEEE Symposium in Low-Power and High-Speed Chips (COOL CHIPS), Tokyo, Japan, 2021, pp. 1-3, doi: 10.1109/COOLCHIPS52128.2021.9410323.
- [3] C. J. Schaefer, P. Taheri, M. Horeni and S. Joshi, "The Hardware Impact of Quantization and Pruning for Weights in Spiking Neural Networks," in IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 70, no. 5, pp. 1789-1793, May 2023, doi: 10.1109/TCSII.2023.3260701.
- [4] Rathi, Nitin & Panda, Priyadarshini & Roy, Kaushik. (2017). STDP Based Pruning of Connections and Weight Quantization in Spiking Neural Networks for Energy Efficient Recognition. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems. PP. 10.1109/TCAD.2018.2819366.

Thank you for your attention!