Optimization of Neuromorphic System Using Approximate Neuron and 3D Stacking Memory

s1290176 Ryoji Kobayashi, Supervisor: Khanh Nam Dang

1. Summary of the Research

Background

Neuromorphic computing (NC) on **3D-IC** has the potential for energy-efficient operations [1]. **SNN** is a low-power solution for AI applications used in the NC system. Since it has noise resilience property**, approximate computing** can be exploited to obtain further energy efficiency. To achieve this, the degree of approximation must be considered carefully.

Research Goal

- Implementation of SNN using approximate neurons and 3D stacking memory.
- Development of a time-efficient optimization algorithm that decides the degree of approximation.
- Evaluation of trade-offs between energy saving and accuracy loss of the SNN applications.

2. Approach/Methodology

2.1 Approximate LIF Neuron

The approximate neuron is realized using the approximate adder that accumulates input value. It outputs a 1-bit spike signal when the accumulated value exceeds the threshold.

2.2 Approximate Stacking Memory

The memory is implemented in the form of stacked dies. Thanks to this architecture, the memory is approximated by employing voltage scaling techniques to different dies.

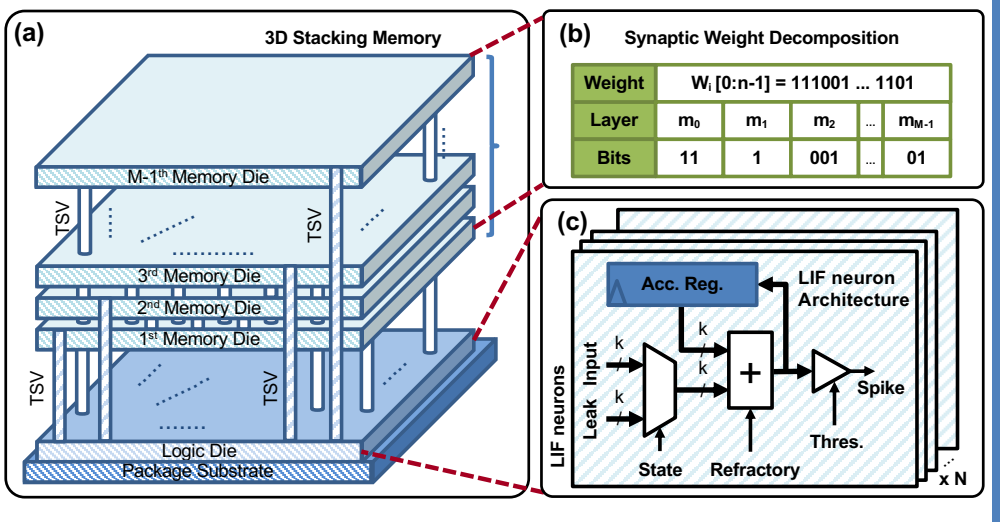

Fig 1. Overview of the hardware architecture.

3. Current Results and Status

SNN is implemented using approximate adders, and an optimization algorithm is developed and evaluated. Here, different adders are applied to different layers of the neural network.

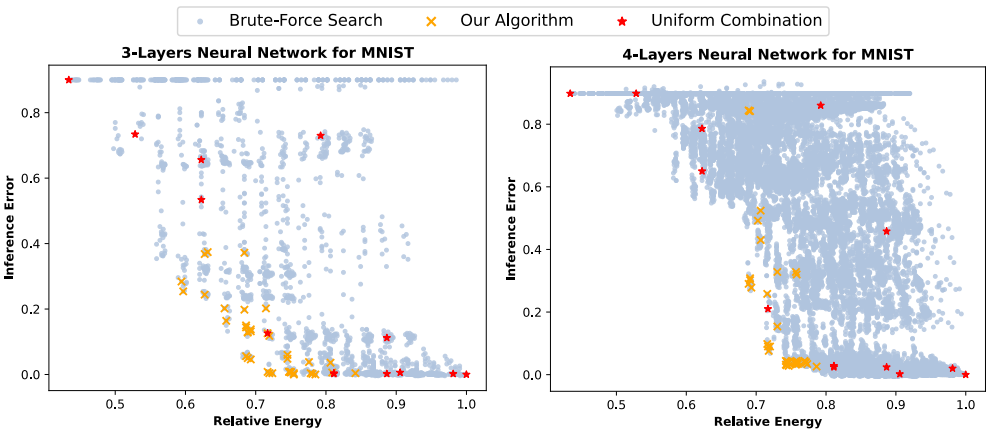

Fig 2. Evaluation of the proposed algorithm.

- Although uniform combinations can achieve energy savings, there is room for further optimization.
- The proposed algorithm can find solutions on pareto-front of the practical region.

Table 1. Number of NN executions

Method	3-layers NN	4-layers NN
Brute-Force	2197	28561
Our Algorithm	~75	~130
Uniform	13	13

- The number of executions is reduced by 96.5% and 99.5% compared to the Brute-Force approach.

4. Remaining Tasks and Schedule

	10	11	12	1
[a]	→	→		
[b]	→	→		
[c]		→	→	→

[a] Evaluate the proposed algorithm with VGG16.
[b] Evaluate the impact of memory approximation.
[c] Writing a thesis

References

[1] N.-D. Nguyen, *et. al.,* "Power-aware neuromorphic architecture with partial voltage scaling 3D stacking synaptic memory," IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 2023.