Architecture and Design of a Spiking Neuron Processor Core Towards the Design of a Large-scale Event-Driven 3D-NoC-based Neuromorphic Processor

Mark Ogbodo,*, *Khanh* Dang, *Fukuchi* Tomohide, and *Abderazek* Abdallah

[1]Adaptive Systems Laboratory, Graduate School of Computer Science and Engineering, The University of Aizu, Japan.

Abstract. Neuromorphic computing tries to model in hardware the biological brain which is adept at operating in a rapid, real-time, parallel, low power, adaptive and fault-tolerant manner within a volume of 2 liters. Leveraging the event driven nature of Spiking Neural Network (SNN), neuromorphic systems have been able to demonstrate low power consumption by power gating sections of the network not driven by an event at any point in time. However, further exploration in this field towards the building of edge application friendly agents and efficient scalable neuromorphic systems with large number of synapses necessitates the building of small-sized low power spiking neuron processor core with efficient neuro-coding scheme and fault tolerance. This paper presents a spiking neuron processor core suitable for an event-driven Three-Dimensional Network on Chip (3D-NoC) SNN based neuromorphic systems. The spiking neuron Processor core houses an array of leaky integrate and fire (LIF) neurons, and utilizes a crossbar memory in modelling the synapses, all within a chip area of 0.12mm^2 and was able to achieves an accuracy of 95.15% on MNIST dataset inference.

1 Introduction

Neuromorphic computing which is aimed at modeling the biological brain on hardware has gone through decades of research [1], and the ability of the biological brain to carryout rapid parallel computations in real time, in a fault tolerant and power efficient manner is the inspiration behind it [2]. The third generation of Artificial Neural Network (ANN) Spiking Neural Network (SNN) has proven to be more effective than its predecessors in this aim, mimicking more closely, the behavior of a biological neuron. The computation of Spiking neurons, like biological neurons are event triggered and communicate via spikes which could be sparse, and this makes them process information only when spikes are received. Neuromorphic architectures take advantage of the sparsity of spikes in SNN to reduce power consumption by power gating parts of the network that are not receiving spikes at any point in time. However, an efficient neuromorphic hardware targeted towards edge application and scalable neuromorphic architecture with large number of synapses requires building small sized neural Processors with low power consumption, efficient neuro-coding scheme, and fault tolerance.

To enable scalability while maintaining minimal power consumption and footprint, we presented in our previous work [3] a Three Dimensional Network-on-Chip (3D-NoC) SNN based architecture, a different approach from the conventional 2D-NoC which is limited in scalability, and consumes more power with increased latency and foot print, when scaling is attempted. The 3D-NoC based SNN architecture utilizes the merits of Network-on-Chips and 3D-Integrated Circuits [4] to enhance the parallelism and scalability of a neuromorphic processor in the third dimension, minimizing power consumption and communication latency as a result of the brief length and low power consumption of the Through Silicon Vias (TSVs) [5] employed in inter-layer communication [6][7]. The 3D-NoC SNN based architecture has the spiking neuron processor cores as the processing elements. These processing elements are connected in a 2D mesh topology to form tiles, and then stacked to form the 3D structure. Communication among the processing elements are made possible with 3D routers [8] (one for each spiking neuron processor core).

In this work, we present the architecture and design of a spiking neuron processor core described in Fig 1 suitable for the 3D-NoC based SNN architecture. The spiking neuron processor core is designed using the leaky integrate and fire (LIF) spiking neuron model which accumulates incoming spikes as membrane potential and stores in the buffer while experiencing leak, then fires an output spike when the membrane potential crosses a threshold. We have chosen the LIF spiking neuron model because of its simplicity, while maintaining some degree of biological plausibility, making it easier to implement. In designing the spiking neuron processor core, we utilized an SRAM for the N×N crossbar based synapse (N is the number of neurons) which has the synapse at the intersection of horizontal and vertical wires that represent the axons and dendrites of the neurons. An SRAM is also used for the neuron and synapse memory. A control unit implemented as a finite state machine is used to control the operations of the spiking neuron processor.

*Corresponding author e-mail: d8211104@u-aizu.ac.jp

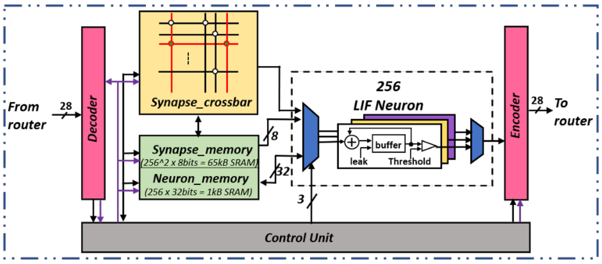

Figure 1. Spiking Neuron Processor Core Architecture.

2 Methodology

The spiking neuron processor core design is described using Verilog-HDL. Cadence tools were used for the synthesis and simulation. The hardware complexity is evaluated for power and area. For performance evaluation, the neuro-core is used to classify MNIST dataset [9] of 60,000 training, and 10,000 inference images on an SNN with an architecture of 748×48×10 trained off-chip with backpropagation as an ANN, then converted to SNN [10]. The MNIST images are converted to spikes using Poisson distribution before being sent to the network for classification. Finally, the result is compared with some existing work and presented in Figure 2.

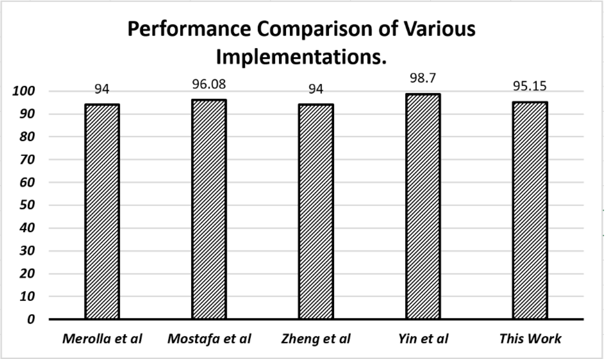

Figure 2. Area and Accuracy comparison

3 Result

The spiking neuron processor consumes an estimated power (leakage and dynamic) of 493.5018mW, covers a chip area of 0.12mm^2 and achieves an accuracy of 95.15% on MNIST dataset inference. The result was compared with some existing works reviewed in [11]. The comparison shows that the spiking neuron processor core has a good trade-off between area and accuracy

4 Conclusion and Future Work

This work presents the architecture and design of a spiking neuron processor core for 3D-NoC SNN, and evaluated its hardware complexity and performance. Future works towards realizing the 3D-NoC SNN architecture will require integrating the spiking neuron processor core into it, and exploring applications that will leverage the architecture.

References

[1] D. Monroe, *Neuromorphic computing gets ready for the (really) big time* (Association for Computing Machinery (ACM), 2014), Vol. 57, pp. 13–15

[2] T. Wunderlich, A.F. Kungl, E. Müller, A. Hartel, Y. Stradmann, S.A. Aamir, A. Grübl, A. Heimbrecht, K. Schreiber, D. Stöckel et al., *Demonstrating Advantages of Neuromorphic Computation: A Pilot Study* (2019), Vol. 13, p. 260, ISSN 1662-453X

[3] T.H. Vu, O.M. Ikechukwu, A. Ben Abdallah, *Fault-Tolerant Spike Routing Algorithm and Architecture for Three Dimensional NoC-Based Neuromorphic Systems* (2019), Vol. 7, pp. 90436–90452, ISSN 2169-3536

[4] K.N. Dang, A. Ben Ahmed, X. Tran, Y. Okuyama, A. Ben Abdallah, *A Comprehensive Reliability Assessment of Fault-Resilient Network-on-Chip Using Analytical Model* (2017), Vol. 25, pp. 3099–3112

[5] K.N. Dang, A.B. Ahmed, Y. Okuyama, B.A. Abderazek, *Scalable design methodology and online algorithm for TSV-cluster defects recovery in highly reliable 3D-NoC systems* (2017), pp. 1–1

[6] K.N. Dang, M.C. Meyer, A.B. Ahmed, A.B. Abdallah, X. Tran, *A non-blocking non-degrading multiple defects link testing method for 3D-Networks-on-Chip* (2020), pp. 1–1

[7] K.N. Dang, A.B. Ahmed, A.B. Abdallah, X. Tran, *TSV-OCT: A Scalable Online Multiple-TSV Defects Localization for Real-Time 3-D-IC Systems* (2019), Vol. 28, pp. 672–685

[8] H.T. Vu, Y. Okuyama, A. Ben Abdallah, *Analytical performance assessment and high-throughput low-latency spike routing algorithm for spiking neural network systems* (2019), Vol. 75

[9] Y. LeCun, *Mnist database of handwritten digits*, http://yann.lecun.com/exdb/mnist/ (2020-03-28)

[10] P.U. Diehl, D. Neil, J. Binas, M. Cook, S. Liu, M. Pfeiffer, *Fast-classifying, high-accuracy spiking deep networks through weight and threshold balancing*, in *(IJCNN)* (2015), pp. 1–8

[11] M. Bouvier, A. Valentian, T. Mesquida, F. Rummens, M. Reyboz, E. Vianello, E. Beigne, *Spiking Neural Networks Hardware Implementations and Challenges: A Survey* (ACM, June 2019), Vol. 15, pp. 22:1–22:35, ISSN 1550-4832