Charles Brunstad

October 24, 2019

Computer Architecture for AI

Further Power Reduction in the TrueNorth Architecture

A plethora of processors have recently been designed to leverage the inherent power-saving nature of neuromorphic architectures. One such chip, TrueNorth, consists of 4096 neurosynaptic cores tiled in a 2-D array and boasts an efficiency of 400 giga-synaptic operations per watt (1). To be fair, however, this figure still falls quite short of that which the brain can achieve, an astonishing 1 peta-synaptic operations at 20 W. Additionally, although the figures are not directly comparable, TrueNorth’s non-neuromorphic peers such as the TPU can execute 92 tera-ops per second while dissipating roughly 40W on the same process (2). Clearly, there still exists some headroom for advancement.

This project aims to evaluate the effects of expanding the TrueNorth architecture’s dimensionality so as to decrease the number of hops required for both intra- and inter-chip spike communication. In particular, it is proposed that each chip’s 64-by-64 mesh grid of cores is reconfigured so as to form a 16-by-16-by-16 torus. As a result, it is expected that both traffic and power use will decrease. Additionally, because spikes arrive at their destination cores asynchronously, it’s possible that the proposed change may enable computation at speeds greater than real-time in small neuromorphic models.

Simplified versions of both the original architecture and that proposed above will be modeled in a pythonic structural simulator. A simple spiking neural network designed to discriminate between a small set of input values will be built on top of the simulator and trained using Spike-Timing-Dependent-Plasticity (3). Metrics such as wire distance traveled, congestion, and spike transit time will be carefully monitored. The two architecture’s performance across these metrics will be compared, and a conclusion will be drawn after additional consideration of disparities in die space available for compute power.

(1) Akopyan et al.: TrueNorth: Design and Tool Flow of a 65mW 1 Million Neuron Programmable Neurosynaptic Chip. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 34 (10), 2015.

(2) Jouppi et al.: In-Datacenter Performance Analysis of a Tensor Processing Unit. 2017.

(3) Chankyu et al.: Training Deep Spiking Convolutional Neural Networks with STDP-Based Unsupervised Pre-training followed by Supervised Fine-Tuning. Frontiers in Neuroscience, 12, 2018.