Four Simulators of the DANNA Neuromorphic Computing Architecture

Adam Disney, James Plank, Mark Dean

TENN LAB
NEUROMORPHIC
ARCHITECTURES. LEARNING. APPLICATIONS.

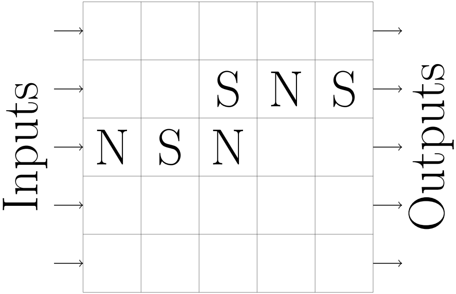
July 25, 2018

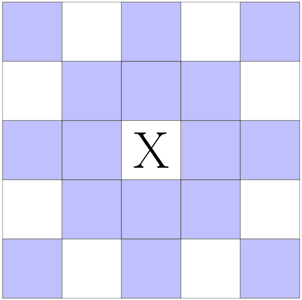
Overview

Introduction

- Moore's Law Ending!
- Digital Spiking Neural Network Hardware
  - IBM TrueNorth
  - Intel Loihi
  - TENNLab DANNA

DANNA Overview

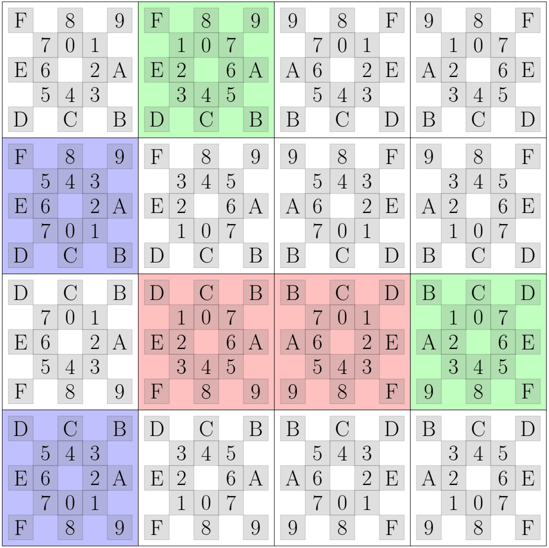




DANNA Quirks

- Two Clocks (Global and Port)
  - Port clock scans neighbors one at a time
  - Neurons can cause additional delay
- Fixed threshold
- Port Orientations enable online learning

Port Orientations



TENN LAB NEUROMORPHIC
ARCHITECTURES. LEARNING. APPLICATIONS.

Goals for DANNA Simulation

- Hardware Verification
- Communication
- Training (EONS)
- Utilization of available computing resources
- Exploration of large devices
- Exploration of future hardware

Simulator #1 - Clock-based Simulator

- Mimic hardware as much as possible
- Hardware I/O
- Small memory footprint
- Array of Unions (Neuron and Synapse)
- Look-up table for ports
- Two loops per port cycle

Simulator #2 - Event-based Simulator

- Central Priority Queue of events eliminates null work
- Converted from polling neighbors to pushing events
- Broadcasting list to determine arrival times
- Used to train NeoN network on all of ORNL Titan for 24 hours

NeoN

https://www.youtube.com/watch?v=d-cjX1n7bqU

Challenges with GPU Programming

- GPUs work best at duplicate tasks on multiple data points
- Divergence is costly
- High arithmetic intensity needed to hide slow global memory
- Can mitigate global memory cost with coalesced access

GPU Simulator - Single Network

- Clock-based simulator
- Divergence due to two element types
- Divergence eliminated by assigning warps to one element type
- Non-contiguous memory access due to port orientations
- Low arithmetic intensity
- Need large networks to get performance

GPU Simulator - Multiple Networks

- Common use case - EONS with small networks
- Each block is an individual simulator
- Why not just run multiple versions of single?
  - Limit on active kernels
  - Can optimize at block level

Stress Test

N	←	N	←
↑	↙	↓	↖
N	→	N	→
↑	↗	↓	↘

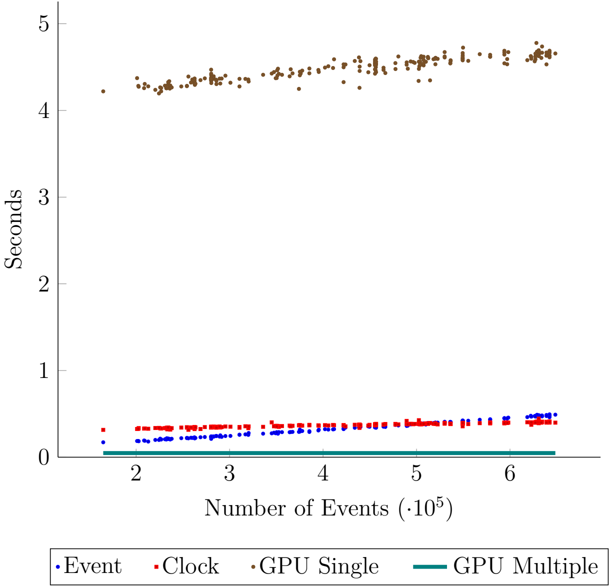
Stress Test

- 224 tests for each network size for 10K cycles
  - 15 x 15
  - 80 x 80
- Randomized input generates random number of total events
- Clock runtime = $\mathcal{O}(tn)$
- Event runtime = $\mathcal{O}(e \log e)$

Test Machine

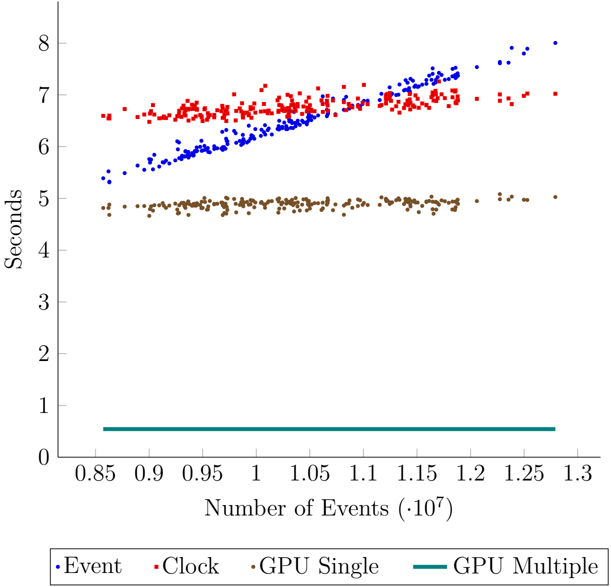
- Dual Xeon E5-2697 v3 @ 2.60GHz
- GeForce GTX TITAN
- Compiled with gcc v5.4.0 and CUDA 9.1

15 by 15 Network



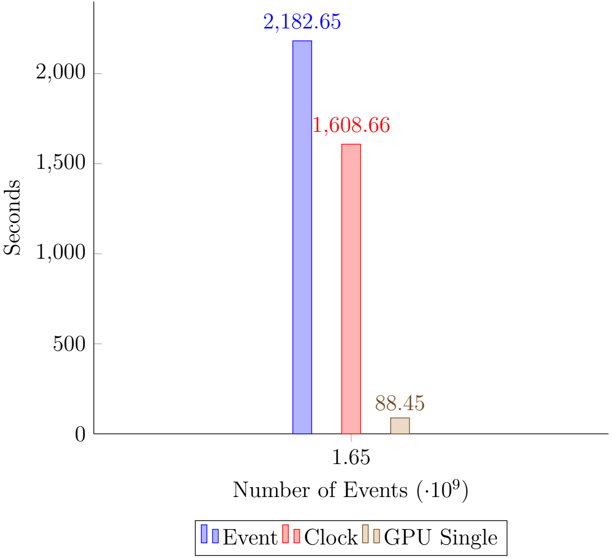
TENN LAB
NEUROMORPHIC
ARCHITECTURES. LEARNING. APPLICATIONS.

80 by 80 Network



TENN LAB
NEUROMORPHIC
ARCHITECTURES. LEARNING. APPLICATIONS.

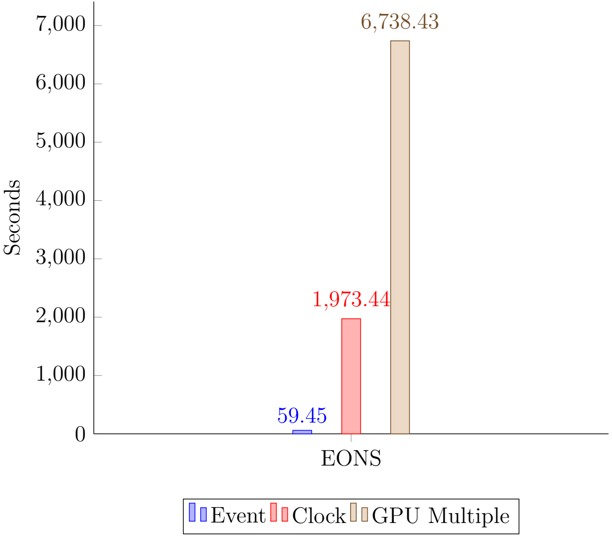
1000 by 1000 Network



EONS Test

- Classification on Breast Cancer Wisconsin dataset from UCI
- Array size = 27 X 27
- Population size of 1000
- 20 epochs where each network runs fitness
- 105K cycles per fitness run
- All EONS runs used the same randomization seed
- Average Neurons = 34.9
- Average Synapses = 61.8
- Average Events per 105K cycles = 26,413

EONS Test Results



Lessons Learned and Conclusions

- Clock-based useful in development of hardware and communications
- Event-based allowed much faster training (NeoN, etc.)
- GPU single network may be useful for future endeavors
- Current use case leaves event-based far superior

Four Simulators of the DANNA Neuromorphic Computing Architecture

Adam Disney, James Plank, Mark Dean

TENN LAB
NEUROMORPHIC
ARCHITECTURES. LEARNING. APPLICATIONS.

July 25, 2018