Do we need high accuracy arithmetic to simulate the brain?

Mantas Mikaitis
mantas.mikaitis@manchester.ac.uk

Supervisor **Dr David R. Lester**
david.r.lester@manchester.ac.uk

Co-Supervisor **Prof. Steve Furber**
steve.furber@manchester.ac.uk

1 Abstract

This question has a big impact on the next-generation neuromorphic machines that are designed to simulate brain activity, specifically for ARM based **SpiNNaker** designed here in Manchester. If the answer is **NO**, then we can build cheaper (in terms of circuit area and leakage power) hardware accelerators, for example integer instead of floating-point arithmetic with narrower datapath and higher latency than ARM cores (32-bit). If the answer is **YES**, then we can aim to find the accuracy required in the arithmetic circuits to simulate specific models. In this poster I demonstrate recent work in this area and discuss the next steps to understand numerical precision required for SpiNNaker-2 - the next generation neuromorphic processor that is currently in design stage.

2 Brain Simulations

Pictured is a typical structure of a neuron. It is estimated that the human brain has 100 billion neurons, each receiving signals from approximately 8000 synapses spiking at average rate of 3Hz and requires only 20W power.

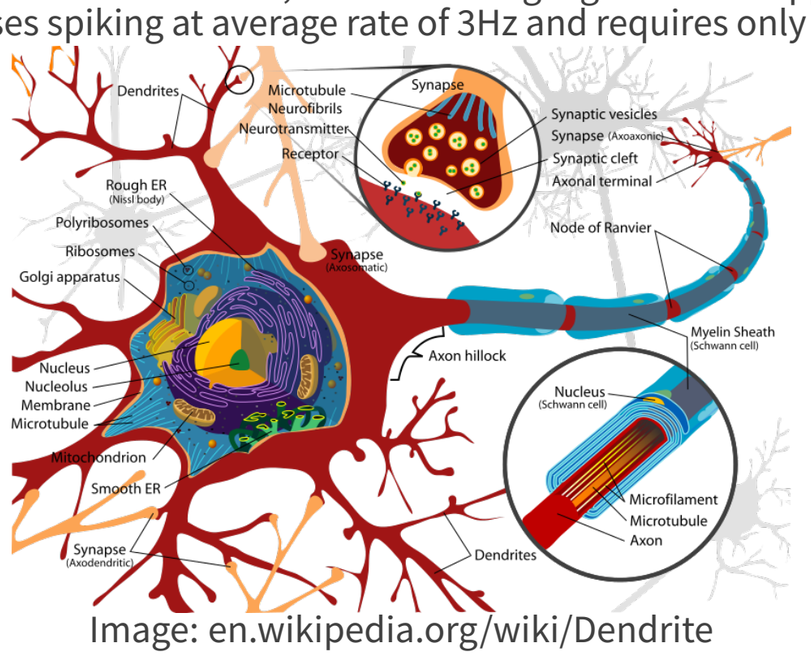
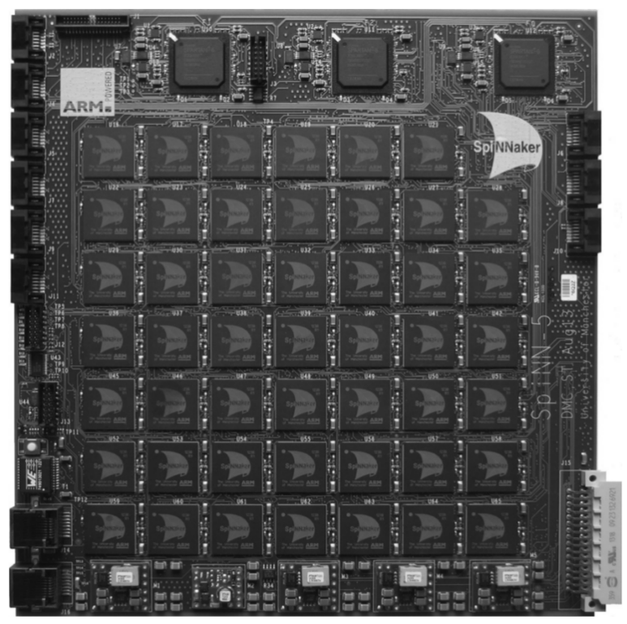


Image: en.wikipedia.org/wiki/Dendrite

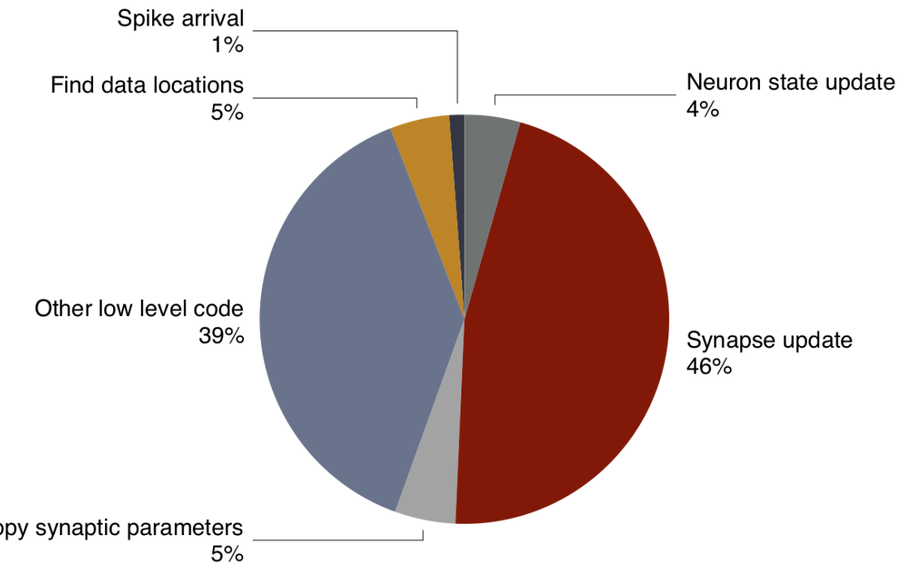
3 SpiNNaker-1 Neuromorphic Computer



- 48 1W chips - each containing 18 classical ARM cores.
- Neural models are mapped into ARM code.
- Neuromorphic part: Signals travel from core to core.
- Signal arriving to a core causes data copying and runs some code to update neural and synapse states.
- Machine with 0.6 million cores is available and can simulate 2 mice brain.
- SpiNNaker-2 is being developed - each chip has 10x computational power @ 1W.
- SpiNNaker-2 has accelerators to run faster than ARM.

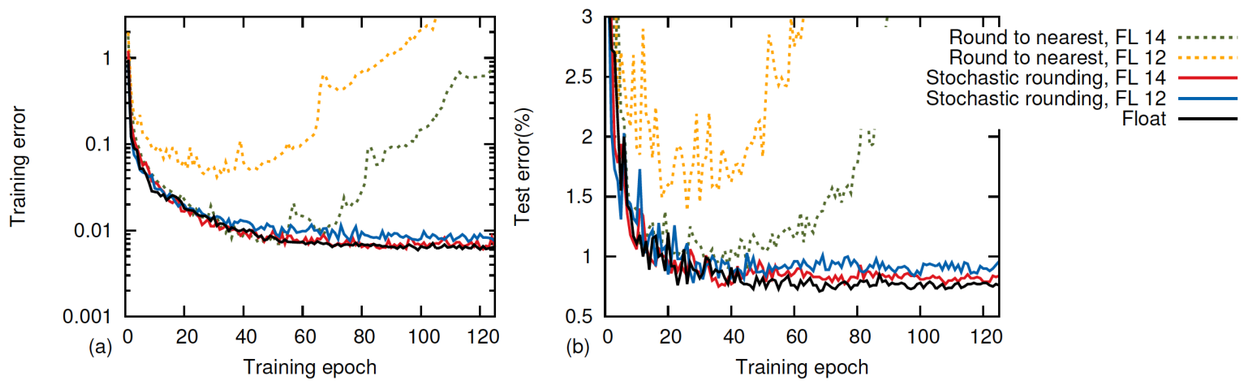
4 Performance of a complex network simulation

Here we have evaluated the resource usage of a single core in the most complex synaptic plasticity rule recently developed on SpiNNaker [1]. 1s simulation of a 40 neurons, targeted by 8000 3Hz spiking neurons with 25% connection density. Additionally, all 40 neurons are receiving 4.5Hz neuromodulator signal. **The most expensive part of simulation, synapse updating, is mainly comprised of 32-bit arithmetic**.



5 Deep Learning with Limited Numerical Precision

Deep learning with limited numerical precision has recently been presented by [2]. Instead of using 32-bit floating point numbers, the authors evaluate fixed point representation with different sizes of fractional bits (FL). Stochastic rounding is used to round the numbers to a given number of fractional bits. It is shown that floating-point representation has negligible advantage over fixed-point on a well known MNIST digit classification benchmark:



16-bit fixed-point add is 18x cheaper in energy and 62x in area than 32-bit floating-point add; 8-bit fixed-point multiply is 5x cheaper in energy and 6x in area than 16-bit floating-point multiply [3].

6 Approximate Computing

Introduce some level of error in the circuit to reduce area, power and delay. Approximate multiplier (built from the 2x2 multipliers) achieves an average power saving of 45% for an average error of 3.3%. By running image sharpening algorithm, the authors show minimal loss in image quality.

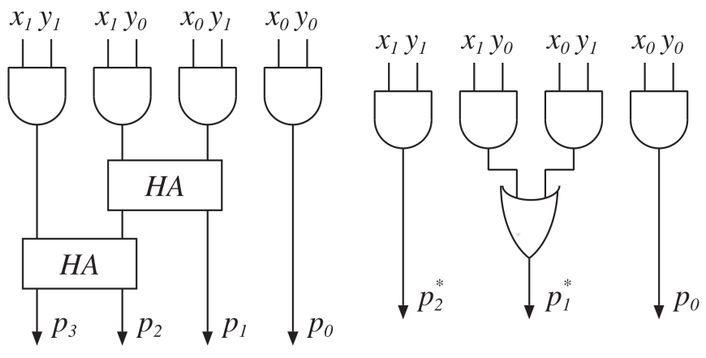
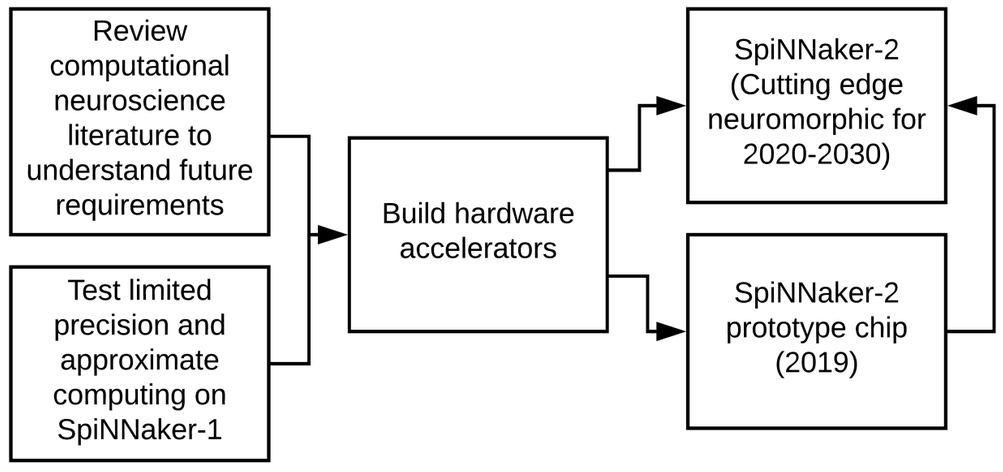


Image: 2x2 multiplier (left) and approximate multiplier (right) with error probability of $\frac{1}{16}$ [4].

7 Planned experiments and SpiNNaker-2 implications

Hypothesis: Brain activity algorithms can tolerate some loss in arithmetic quality caused by limited arithmetic precision and approximate circuits.



References

[1] Mantas Mikaitis, Garibaldi Pineda García, James C. Knight, and Steve B. Furber. Neuromodulated Synaptic Plasticity on the SpiNNaker Neuromorphic System. *Frontiers in Neuroscience*, 12:105, 2018.

[2] Suyog Gupta, Ankur Agrawal, Kailash Gopalakrishnan, and Pritish Narayanan. Deep Learning with Limited Numerical Precision. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 1737–1746. JMLR.org, 2015.

[3] William Dally. High-Performance Hardware for Machine Learning, U.C. Berkeley October 19. 2016.

[4] Miloš D. Ercegovac. On approximate arithmetic. *Conference Record - Asilomar Conference on Signals, Systems and Computers*, pages 126–130, 2013.