Referee Report

Paper Number:
Reviewer's Name: Jesse Bannon
Name of Paper: TrueNorth Ecosystem for Brain-Inspired Computing: Scalable Systems, Software, and Applications
Author(s): Jun Sawada, ..., (IBM Research, Lawrence Livermore National Laboratory, US Air Force Research Laboratory, US Army Research Laboratory)

Section I. Overview

A. Reader Interest

1. Which category describes this manuscript?

( ) Practice / Application / Case Study / Experience Report
(x) Research / Technology
( ) Survey / Tutorial / How-To

2. How relevant is this manuscript to the readers of this periodical? Please explain your rating under IIIA.

( ) Very Relevant
( ) Relevant
(x) Interesting - but not very relevant
( ) Irrelevant

B. Content

1. Please explain how this manuscript advances this field of research and / or contributes something new to the literature. Please explain your answer under IIIA. Public Comments.

2. Is the manuscript technically sound? Please explain your answer under IIIA. Public Comments.

( ) Yes
(x) Appears to be - but didn't check completely
( ) Partially
( ) No

C. Presentation

1. Are the title, abstract, and keywords appropriate? Please explain your answer under IIIA. Public Comments.

(x) Yes
( ) No

2. Does the manuscript contain sufficient and appropriate references? Please explain your answer under IIIA.

(x) References are sufficient and appropriate
( ) Important references are missing; more references are needed
( ) Number of references are excessive

3. Does the introduction state the objectives of the manuscript in terms that encourage the reader to read on? Please explain your answer under IIIA. Public Comments.

(x) Yes
( ) Could be improved
( ) No

4. How would you rate the organization of the manuscript? Is it focused? Is the length appropriate for the topic? Please explain your answer under IIIA. Public Comments.

(x) Satisfactory
( ) Could be improved
( ) Poor

5. Please rate and comment on the readability of this manuscript. Please explain your answer under IIIA.

(x) Easy to read
( ) Readable - but requires some effort to understand
( ) Difficult to read and understand
( ) Unreadable

Section II. Summary and Recommendation

A. Evaluation

Please rate the manuscript. Please explain your answer under IIIA. Public Comments.

( ) Award Quality
(x) Excellent
( ) Good
( ) Fair
( ) Poor

B. Recommendation

Please make your recommendation. Please explain your answer under IIIA. Public Comments.

( ) Accept with no changes
(x) Author should prepare a minor revision
( ) Author should prepare a major revision for a second review
( ) Reject

Section III. Detailed Comments

A. Public Comments (these will be made available to the author)

Explanation for the Recommendation

The TrueNorth ecosystem showcases an exciting technology that is capable of executing sophisticated deep neural networks at a fraction of what modern computing devices require for energy. The research presented in this paper contributes significantly to HPC's overall objective of reducing energy consumption. It is well written and informative, and deserves a rating of excellent.

The paper is well organized start-to-finish. Every graphic, especially Fig. 1, reinforce the contents of the paper. Describing the TrueNorth ecosystem bottom-up is especially helpful when describing the entire architecture stack.

The only revision that I recommend is in the introduction: “On the other hand, neurosynaptic architectures, such as TrueNorth, have demonstrated orders of magnitude improvement in computational energy-efficiency and throughput”. In the related works section, the only metric comparing TrueNorth to other systems is FPS/W. While this metric does prove the energy-efficiency, it does not suffice for a throughput metric. If TrueNorth can demonstrate an order of magnitude improvement in throughput, data should back this up. If not, the research is still very significant, and I would recommend to change just that one sentence.

Summary of the Paper and Assessment

The TrueNorth Ecosystem aims for power efficiency and throughput in deep learning. Execution is composed by the TrueNorth architecture, which spans the entire hardware/software stack. The paper presents the entire stack bottom-up, followed by an evaluation of convolutional neural networks characterized by frames/sec/Watts (FPS/W). The objectives are clear and organization is stated and displayed by Fig. 1, making the contents of the paper clear to the reader.

The TrueNorth architecture does not take the Von-Neumann approach. Its constructed similar to a multi-layer perceptron: each 'core' represents a fully-connected neural network with 256 input/output neurons connected by 256-by-256 synapses. Each chip, named NS1e, contains 4096 cores tiled as a 64-by-64 array. The cores are built to be composable, allowing connectivity between many cores to expand the depth of a network.

TrueNorth is able to both scale-out and scale-up. The NS1e-16 platform scales out cores by connecting them via commodity networking hardware. They demo this feature connecting sixteen NS1e boards together with the option to include a quad-core Xeon to act as a host gateway within a single 6U rack, which consumes a total of 68W not including the server.

The NS16e platform scales up by integrating sixteen NS1e chips onto a single board. This significantly decreases the latency of communication between chips. Similarly to NS1e-16, NS16e can be coupled with traditional processors such as a CPU/GPU/FPGA for pre/post processing.

Executing code on this architecture requires users composing their networks in Corelet Programming Language (CDL), a MATLAB based suite capable of mapping software onto the TrueNorth architecture. The authors do not go into much detail of this language or usability. They instead focus on the mapping from software to the cores; how weights are clustered into groups to minimize communication. This decision was appropriate because the focus of this paper was throughput and energy efficiency, which is derived from TrueNorth's hardware.

Multiple pages are dedicated to describing use cases of the TrueNorth ecosystem. Each example showcases TrueNorth being used for image recognition applications. Despite this section not providing more details about TrueNorth itself, it informs the reader that the entire stack of the TrueNorth ecosystem is usable and production-ready.

Lastly, performance evaluations show TrueNorth is significantly more efficient than every system capable of executing neural networks, including SpiNNaker, a similar ecosystem with dedicated hardware for neural nets. It is not clear if their claim of “orders of magnitude improvement in throughput” is true, because they only consider FPS/W. The authors should revise that single sentence in their introduction, or provide the performance results to prove that claim.