Final Technical Report

DE-SC0021335
A Hardware and Software Co-design Framework for Energy Efficient Neuromorphic Systems

Hai Li, Duke University

Accomplishments:

1. What are the major goals of the project?
    a. In the first 9- months of the project, the major goals include investigating:
        i. The sparsification potential for large- scale computing applications
        ii. Structured sparsity learning techniques for sparse weights in data parallelism
        iii. Unsupervised learning and initialization strategies
    b. Overall, in the 2- year performance period, we want to evaluate our algorithm improvement with deployment constraints. Final deliverables will include:
        i. A circuit implementation of an RRAM- based temporal coding SNN PE and design methods to enhance the robustness
        ii. An asynchronous architectural framework for the acceleration of inference and online training
        iii. Algorithm- level techniques that minimize communication cost for higher system- level efficiency
        iv. A report detailing the study and recommendations across circuit implementation, architecture, framework, algorithm structure, experimental setup, and evaluation results.
2. What was accomplished under these goals?
    a. Major Activities
        i. Implementing synapses with ReRAM devices has been widely investigated, whereas most works focus on realizing either spike-timing-dependent plasticity (STDP) or supervised learning rules with ReRAM devices [Hu 2016][Lashkare 2017][Shrestha 2017]. Few concentrate on the circuit implementation of the efficient inference of the event-driven SNN models, especially those leveraging temporal patterns of spikes. Complex event-driven neuron dynamics was an obstacle to implementing efficient brain-inspired computing architectures with VLSI circuits. To solve this problem and harness the event-driven advantage, we propose ASTERS, a resistive random-access memory (ReRAM) based neuromorphic design to conduct the time-to-first-spike SNN inference. In addition to the fundamental novel axon and neuron circuits, we also propose two techniques through hardware-software co-design: “Multi-Level Firing Threshold Adjustment” to mitigate the impact of ReRAM

device process variations, and “Timing Threshold Adjustment” to further speed up the computation.

ii. On-chip learning is often performed through back-propagation, which is nonoptimal for execution on most hardware platforms as it requires transposed copies of each layer’s weights and sequential operations on errors propagating from output layers to input layers. Popular alternatives such as spike-timing-dependent plasticity (STDP) often result in lower accuracy models. This has led to the development of novel algorithms like equilibrium propagation and its spiking descendents, in which the changing states of a recurrent neural network are used to update the weights until convergence. Equilibrium propagation and spiking equilibrium propagation[Scellier and Bengio 2017][Martin et al 2021] can be made more efficient through analog computing schemes. To this end, we develop an analog neuron design in CMOS technology which computes the spike rate activation and its derivative for a neuron to directly update analog synapses in real-time. Changes to the algorithm, such as ternary spike rate activation gradients, are introduced to simplify hardware implementation.

b. Specific Objectives

i. The main objectives of this project include: 1) developing SNN algorithms that improve the energy efficiency of a SNN through high- sparsity data representation, 2) co- designing SNN processing elements that implement these algorithms and are capable of both online training and inference, and 3) integrating these processing elements into a cohesive and asynchronous architecture. In a spiking neural network, the aim of high- sparsity data is to reduce the number of voltage spikes representing a sample of data, thus reducing the energy consumption of the system. Temporal- coding represents a single datum as the timing of a spike relative to some reference (such as other spikes), and therefore needs fewer spikes than a rate- coding system where the frequency of many spikes represents the datum. Coupled with other methods of introducing sparsity, such as winner- takes- all systems where only a single neuron in a layer spikes, can further reduce the number of spikes and subsequent energy consumption of a system. When these algorithms are co- designed with the processing element circuits, emerging technologies and novel circuit techniques can be utilized to optimally implement the algorithms with fewer components. A device’s properties can perhaps be leveraged to emulate a portion of the algorithm physically, requiring fewer devices than if the algorithm was digitally executed. The efficiency of the system can be further optimized one level higher by the architecture that tiles and integrates the processing elements. An asynchronous connection of PEs would require no global clock and would reduce time spent waiting by data moving from element to element. Furthermore,

temporal- coded data needs to travel atomically to preserve spike- times, which could be complicated by synchronized interconnects. All three of these objectives serve to improve the energy and area efficiency of a neuromorphic processor without sacrificing accuracy or performance.

c. Significant Results
    i. We use a twin-column excitatory/inhibitory synaptic weight mapping scheme and design a set of spike-timing axon and neuron circuits for ReRAM crossbars. Regarding the widely-concerning process variations of ReRAM nano-devices, we propose a multi-level “firing threshold adjustment” method that effectively recovers the inference accuracy degradation with only 1.5% hardware overhead. We also devise the “timing threshold adjustment” method, a design-automation technique specifically for the proposed postsynaptic neuron circuits. It removes a considerable amount of unnecessary spike generation between adjacent neural network layers and thus speeds up the inference. Experimental results show that our cross-layer solution achieves more than 34.7% energy savings compared to the existing spiking neuromorphic designs, meanwhile maintaining 90.1% accuracy under the process variations with a 20% standard deviation.
    ii. For implementation of spiking equilibrium propagation in real-time, synapse weights need to be updated often and without stopping the sensing function to update the weight. Therefore, we utilized a CMOS synapse [S. Kim et al 2017, Y. Li et al 2018] which can update often without significantly degrading performance and can include an extra port for weight updates without disconnecting from the output IFC. To ensure linear weight updates, we assumed ternary gradients, which did not greatly affect classification accuracy for (a purely software) simulation of a 1-layer network trained over five epochs on five MNIST [L. Deng 2012] classes (97.49% vs. 96.49% for full-precision). The analog neuron and synapse implementations were functionally correct, increasing synapse weight as spike rate increased and decreasing synapse weight as spike rate decreased. The average power consumed by the neuron was 95.75uW, with only 79.79nW pulled by the synapse. Future research in this area will investigate incorporation of new ReRAM-based synapse structures for equilibrium propagation and decreasing neuron power consumption through optimization and mixed-signal techniques.

d. Key Outcomes or Other Achievements
    i. Demonstrated functionality and evaluated performance of the proposed ReRAM-based neuromorphic design. Analysis of parameter set-up of the proposed threshold adjustment schemes.
    ii. Demonstrated functionality of analog CMOS implementation of promising new algorithm for online and on-chip training of spiking neural networks.

3. What opportunities for training and professional development has the project provided?
    a. In this reporting period, student researchers have had the opportunity to publish papers on the topic of temporally-coded SNNs, particularly in the areas of efficient, time-to-first-spike-based inference with ReRAM circuits and on-chip implementation of promising new learning algorithms. These publications are currently being prepared for presentation at IEEE conferences during the summer.
4. How have the results been disseminated to communities of interest?
    a. As mentioned, the research outcomes for this reporting period have been accepted for publication in conference proceedings and presentation at their respective conferences. These conferences include the 2022 Design Automation Conference and the 2022 IEEE International Conference on Artificial Intelligence Circuits and Systems.
5. What do you plan to do during the next reporting period to accomplish the goals?
    a. Our areas of focus in the next reporting period include:
        i. Complete neuromorphic vision system designs that incorporate the SNN processing element with sensor frontends. Our research work will target at the design of neuromorphic in-sensor-processing vision systems that can pre-process the sensory data within sensor pixels in the spike format. By deploying digital or analog processing circuits close to the photodetectors, the sensed images can be processed locally without costly data transmission and conversion. Novel intra-pixel processing circuits and inter-pixel communication schemes should be explored to directly perform basic multiply-and-accumulate computation with photocurrents.
        ii. New RRAM-based synaptic structures for efficient implementation of emerging local learning rules like equilibrium propagation for SNNs. Our research work has demonstrated the functionality of an analog implementation of equilibrium propagation for a CMOS synapse. By developing new synaptic structures featuring RRAM devices, we can potentially retain the same functionality of the CMOS synapse that allows the weight to be updated without stopping sensing functions but with less consumption of area and power as well as longer retention times. New RRAM-based synaptic structures and topologies should be investigated for utilization in local learning rules that update weights through direct interaction of presynaptic and postsynaptic signals.

Publication Details:
Ziru Li, Qilin Zheng, Bonan Yan, Ru Huang, Bing Li and Yiran Chen, “ASTERS: Adaptable Threshold Spike-timing Neuromorphic Design with Twin-Column ReRAM Synapses,” 2022 59th ACM/IEEE Design Automation Conference (DAC). IEEE, 2022.

Brady Taylor, Nicky Ramos, Eric Yeats, and Hai Li, “CMOS Implementation of Spiking Equilibrium Propagation for Real-Time Learning,” *2022 IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS)*, Accepted, 2022.

Intellectual Property Details:
There are no intellectual properties to report.

Technologies and Technique Details:
There are no technologies or techniques to report.

Other Product Details:
There are no other products to report.

Participant Details:
Hai Li
Ziru Li
Brady Taylor

Partner Details:
There are no partners to report.

Other Collaborator Details:
There are no other collaborators to report.

Impact:

1. What is the impact on the development of the principal discipline(s) of the project?
    a. The algorithms and hardware produced by this project will reduce the energy and area footprint of next- generation neuromorphic processors. With online learning capabilities, these designs could prove invaluable for edge computing and internet- of- things applications, where a deployed processor would be able to process large amounts of sensor data in real time and adapt to environmental changes. In general, these algorithms and processors would be practical for any application working on large datasets with strict power and area budgets, such as data centers and wearable electronics. By focusing on algorithms and hardware with sparse data representation and temporal- coding, we are emulating the brain’s ability to process time- series data with minimum energy requirements. Furthermore, by leveraging emerging technologies, we are exploring potential alternatives to the plateauing performance of CMOS von Neumann machines.
2. What is the impact on other disciplines?
    a. Beyond the primary discipline of neuromorphic and high- performance computing of this project, we anticipate that our designs can create a platform for processing and learning features of large amounts of data quickly and efficiently. This could potentially lower the barrier for projects in disciplines where the cost for processing data is prohibitive. Furthermore, we expect that our investigation into

online, in- situ, and biologically- plausible learning in SNNs could potentially shine light on the underlying neuroscience and how neural systems are able to compute and learn so efficiently.

3. What is the impact on the development of human resources?
    a. Integration of research with education will be accomplished through motivation and improved training of undergraduate and graduate students. The expected educational contribution of this work is the establishment of an interdisciplinary hands- on, research- based curriculum for high- performance and energy-efficient manycore chip design that will increase the number of students attracted to this field. Students working with the PIs will be assets to various high- tech companies and academic institutions by virtue of the training they are receiving.
    b. At Duke University, the PI actively recruits women students. She served as Faculty Champion for the Sloan University Center of Exemplary Mentoring (UCEM) at Duke University. The Duke UCEM seeks to increase the number of outstanding Ph.D. graduates from underrepresented populations (i.e., African Americans, Hispanics, and Native Americans) in the physical sciences and engineering. PI Li has been actively involved into the recruiting and mentoring activities, including meeting with potential and current students, attending panels and workshops, etc.
    c. Together with Dr. Catherine Schuman at ORNL, Dr. Mondira Pant at Intel Corp., Professor Duygu Kuzum at UCSD, and Professor Dhireesha Kudithipudi at UTSA, the duke PI Li proposed a panel on “Brain- Inspired Computing for the Edge” at Grace Hopper Conference. The panel was held in September 2020.
4. What is the impact on physical, institutional, and information resources that form infrastructure?
    a. N/A
5. What is the impact on technology transfer?
    a. N/A
6. What is the impact on society beyond science and technology?
    a. N/A
7. Foreign Spending
    a. Not provided

Changes-Problems:

1. Changes in approach and reasons for change.
2. Actual or anticipated problems or delays and actions or plans to resolve them.
3. Changes that have a significant impact on expenditures.
4. Significant changes in use or care of human subjects, vertebrate animals, and/or biohazards.
5. Change of primary performance site location from that originally proposed.
6. Carryover amount.