DOI: 10.22661/AAPPSBL.2020.30.2.21

Neuromorphic Computing for Machine Learning Acceleration Based on Spiking Neural Network

SANGBUM KIM
DEPARTMENT OF MATERIALS SCIENCE AND ENGINEEIRNG, SEOUL NATIONAL UNIVERSITY

ABSTRACT

Recent breakthroughs in deep learning have spurred interest in novel computing architectures that can accelerate machine learning algorithms. Neuromorphic computing utilizing fully parallel operation enabled by an array of resistive elements is being actively explored to implement power- and area-efficient machine learning accelerators. To successfully develop a neuromorphic processor, it is crucial to co-optimize the device, circuit, architecture, and algorithm. In this regard, this article introduces a fully integrated neuromorphic processor using phase change memory as a synaptic device that implements a fully asynchronous and parallel operation of a spiking neural network running a restricted Boltzmann machine as a learning algorithm.

INTRODUNTION

Recent breakthroughs in deep learning have enabled its wide adoption in various disciplines such as real-time language translation, speech recognition, and disease diagnosis. As deep learning is applied to more complex problems and optimized for better accuracy, the number of parameters in deep neural networks is increasing exponentially at a striking pace of 2 times per 3.5 months [1]. On the other hand, the rate of progress in computing power has slowed down significantly as Moore's law is approaching its limit. This widening gap between supply and demand for computing power has spurred interest in new computing paradigms that can potentially provide orders of magnitude improvement in computation for deep learning or machine learning in general. One such candidate is neuromorphic computing using nonvolatile memories [2].

In this article, I will introduce the basic ideas behind neuromorphic computing and the key requirements to realize its true potential. Then, I will introduce a neuromorphic processor that was recently developed using phase change memory to implement a machine learning algorithm on a spiking neural network [3].

NEUROMORPHIC COMPUTING

The concept of neuromorphic computing was originally developed by Carver Mead who described the use of analog electronic circuits to mimic biological information-processing systems [4]. This principle has been widely applied to various areas such as biologically plausible sensors and processors developed to facilitate the study of the biological brain itself. Since deep learning or machine learning in general has delivered astonishing breakthroughs, various ideas have been proposed to accelerate machine learning using neuromorphic computing. This article will focus on the recent trend where a novel computing architecture based on an array of resistive elements is specifically designed to accelerate machine learning algorithms such as deep learning and the like.

Computation Using an Array of Resistive Elements

Deep learning and similar machine learning algorithms can be executed more efficiently using an array of resistive elements. During deep learning, the majority amount of computation time is spent on multiplication of matrixes that store a set of synaptic weights or strengths of connections between neurons in two adjacent layers. For an n-by-n matrix, the computation time for matrix

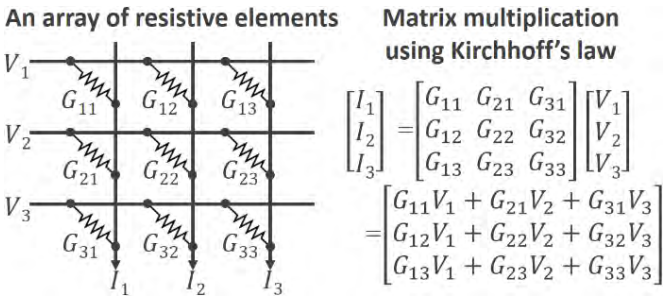

Fig. 1: By applying voltages to an array of resistive elements while measuring the currents, matrix multiplication can be implemented in a fully parallel fashion.

multiplication in incumbent digital computing systems is proportional to n^2. Using an array of resistive elements shown in Fig. 1, the same matrix multiplication can be executed in a fully parallel fashion and the computation time no longer depends on the size of the matrix, which leads to greatly reduced computation time. In addition, computation energy is greatly saved because energy-intensive operations to move elements in matrixes from memory to digital multipliers are no longer needed. Lastly, it can be more area-efficient because each resistive element stores a matrix element in an analog fashion, which can replace multiple digital memory cells.

Various Challenges

To realize the aforementioned benefits of analog computing based on an array of resistive elements for deep learning, there are various practical engineering challenges. First of all, the characteristics of resistive elements storing analog values need to be improved to minimize the accuracy loss caused by analog computing [5]. For example, resistance instabilities such as noise and drift could cause inaccuracies in the computation results. More importantly, the so-called linearity and symmetry of weight update [6] is a key requirement to enable a fully parallel synaptic weight update that can reduce the computation time for training the deep neural network. Currently, a variety of novel materials and devices are being explored as resistive elements based on diverse working principles such as phase change, conductive filament formation, charge storage, and ionic storages [7]. At the same time, various algorithmic and circuit-level remedies are being explored to mitigate the problems caused by non-ideal characteristics of resistive elements. Secondly, the reconfigurability of neuromorphic processors needs to be addressed to apply them to various neural networks of different sizes. For example, a neural network can be split into two or more sub neural networks to map it to a set of arrays of resistive elements [8].

Spiking Neural Network

The same array of resistive elements can be applied to implement a spiking neural network. Various types of spiking neural network have been proposed. One of the most simple spiking neural networks uses leaky integrate and fire (LIF) and spike-timing dependent plasticity (STDP) neurons.

In LIF, each neuron in the spiking neural network can generate a spike (an electrical current pulse with short pulse width and fixed amplitude) asynchronously with respect to other neurons. The generated spike from a firing neuron is delivered to other neurons if they are connected to the firing neuron via a synapse. The weight of a synapse determines the amplitude of the spike thereby representing the strength of the connection between two neurons. When a neuron receives spikes from other neurons, they are capacitively integrated in the form of the membrane voltage. In addition to spikes, a leak current is also integrated, which lowers the membrane potential continuously. When the membrane voltage exceeds a specified threshold, the neuron will generate a spike that will be delivered to other connected neurons and repeat the LIF mechanism described above. After generating a spike, the membrane potential of the neuron is reset to its initial value and the neuron becomes unresponsive for a given refractory time. Biological synapses deliver spikes only in one direction. However, in certain types of spiking neural networks, synapses are allowed to deliver spikes in both directions including the neuromorphic processor introduced in this article.

STDP is a synaptic weight update rule. The amount and timing of the synaptic update is determined solely by the timing of spikes from two neurons which are connected through the synapse. The key characteristic of STDP is that it is a local update rule because the only information it uses for the update comes from two adjacent neurons.

Spiking neural networks are being actively explored by academic and industrial research groups as a way to enable the next generation of AI beyond deep learning [9-11]. However, various key ingredients of spiking neural networks such as learning algorithms, overall architecture, and target application are still missing. Therefore, it is important to co-optimize various aspect of spiking neural networks. In the following section, we will discuss one such example.

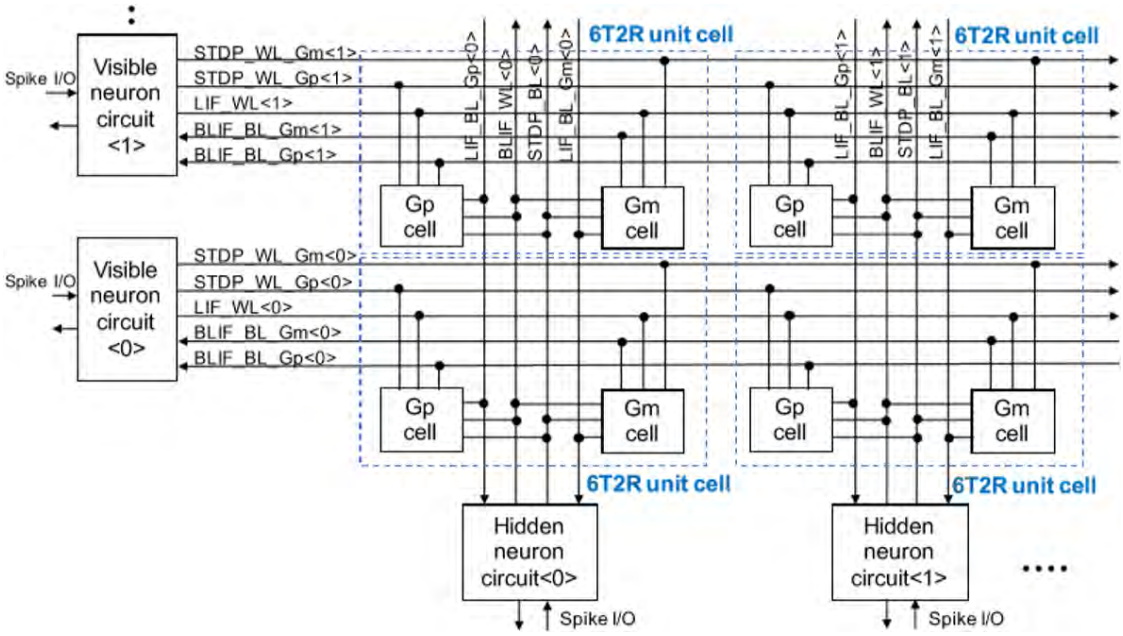

Fig. 2: Overall architecture of a neuromorphic processor. An array of 6T2R synaptic unit cells is surrounded by visible and hidden neuron circuits on the periphery. (Reprinted, with permission, from [3].)

ON-CHIP TRAINABLE SPIKING NEURAL NETWORK NEUROMORPHIC PROCESSOR

Learning Algorithm

To implement a spiking neural network neuromorphic processor especially with on-chip training capability, a governing learning algorithm needs to be determined before designing the corresponding architecture and circuits. One machine learning algorithm that can be used as a learning rule for spiking neural networks while maintaining key features of spiking neural networks such as asynchronous and parallel operation, local update rule, and sparse spike representation is the restricted Boltzmann machine (RBM) [12]. The overall architecture in Fig. 2 is designed such that it can execute three operations comprising RBM in an asynchronous and parallel fashion using the two mechanisms of spiking neural networks previously described as LIF and STDP. Three operations in RBM are forward pass, backward pass, and contrastive divergence weight update.

Phase Change Memory as a Resistive Element

We implemented a resistive element using phase change memory (PCM), which is one of the most mature among various emerging non-volatile semiconductor memory technologies as can be seen from a stand-alone phase change memory product commercialized by leading semiconductor companies [13]. A PCM cell stores the analog synaptic weight in the form of electrical resistance, which is determined by amorphous/crystalline phase distribution. The phase distribution in the PCM cell can be gradually modulated by partially crystallizing the phase change material [14].

Non-ideal characteristics of a PCM as a synaptic device have been mitigated by various measures in the neuromorphic processor. First of all, the asymmetry of PCM weight update has been compensated by adjusting the ratio between RESET and SET programing probabilities. Secondly, the impact of noise on PCM weight is masked by the stochastic behavior of the RBM neuron.

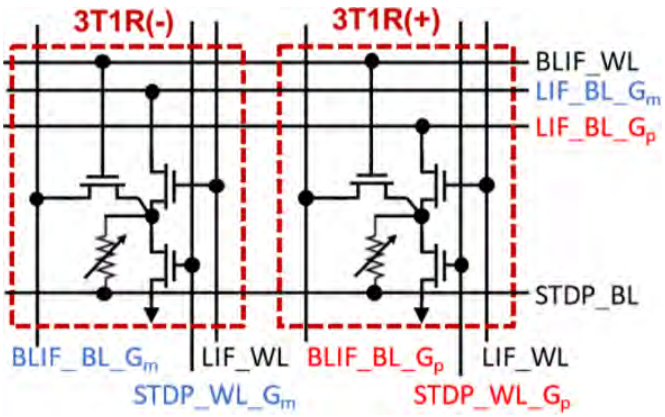

Fig. 3: 6T2R synaptic unit cell comprising two 3T1R half synaptic unit cells. (Reprinted, with permission, from [3].)

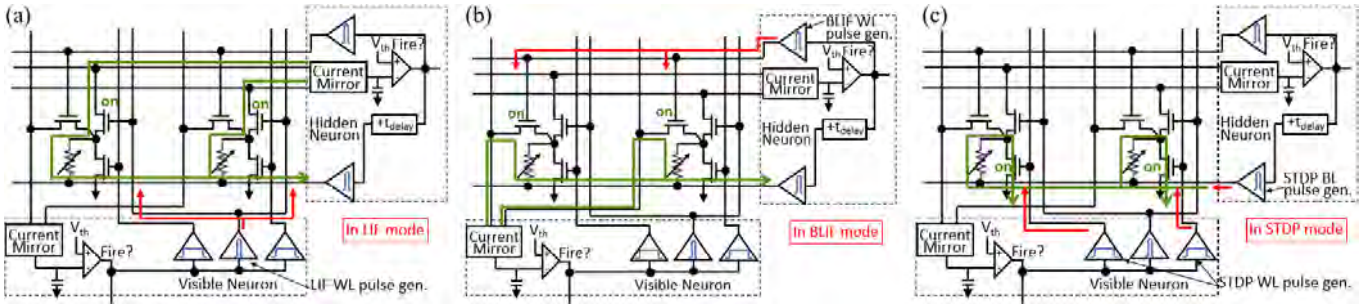

Fig. 4: Asynchronous and parallel operation of a neuromorphic processor implementing (a) forward pass, (b) backward pass, and (c) contrastive divergence for a restricted Boltzmann machine. (Reprinted, with permission, from [3].)

6T2R Synaptic Unit Cell

In this section, a detailed explanation of how we designed the 6T2R synaptic unit cell in Fig. 3 will be provided. We used two identical synaptic half unit cells (3T1R) to comprise one synaptic unit cell (6T2R). The synaptic weight is represented by the difference between 2R's (resistive elements) enabling both positive and negative weights. In each synaptic half unit cell (3T1R), three transistors (3T) are connected to a resistive element (1R) in a way to enable aforementioned 3 key operations of a RBM. The parallel configuration of 3 transistors enables a fully asynchronous and parallel execution of 3 key operations in the RBM.

Visible and Hidden Neuron Circuits on the Periphery

On the bottom side of the array of 6T2R synaptic cells is the visible neuron circuit. The visible neuron circuit generates a forward LIF (leaky integrate and fire) pulse which executes the forward pass in the RBM. On the right side of the array is the hidden neuron circuit. The hidden neuron circuit generates a backward LIF pulse which executes the backward pass in the RBM. The weight update rule of the RBM, contrastive divergence, is implemented by properly designing programming STDP pulsing schemes in both visible and hidden neuron circuits. The synaptic weights are updated only if two pulses from the visible neuron circuit and the hidden neuron circuit arrive at the PCM cell at the same time to induce a large amount of programming current through the PCM cell. The overall operation of the circuits is described in Fig. 4.

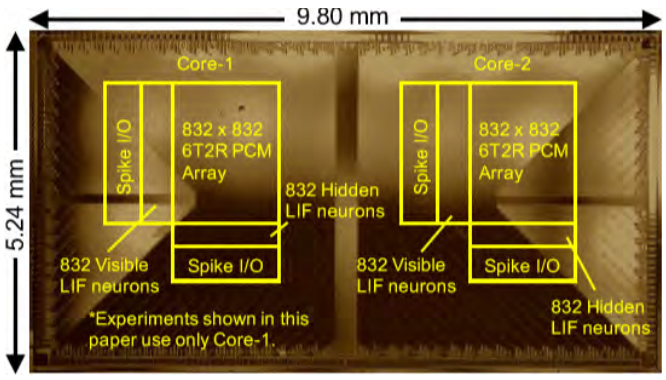

Fig. 5: A fully integrated neuromorphic processor (Reprinted, with permission, from [3].)

For both forward LIF and backward LIF operations, random walk circuits are added to implement the stochastic nature of RBM neurons. The random walk neuron periodically adds or subtracts a fixed amount of voltage from the membrane potential leading to a firing probability exponentially dependent on the membrane potential.

On-Chip Learning and Inference Demonstration

Using the fully integrated chip in Fig. 5, we have demonstrated the on-chip learning and inference capability of a RBM spiking neural network using the MNIST database.

CONCLUSION

In this article, a neuromorphic core implementing spiking neural network that runs a restricted Boltzmann machine is introduced. The neuromorphic core demonstrates that an array of resistive elements can be successfully applied not only to matrix multiplications but also to implementation of a synaptic network in a spiking neural network. Phase change memory has been selected as a synaptic device, the non-ideal characteristics of which have been mitigated by algorithm and circuit implementations. A restricted Boltzmann machine was chosen as a learning algorithm due to its compatibility with a spiking neural network and the feasibility of efficient hardware implementation. Various components in the architecture, such as the synaptic device and the synaptic

unit cell to neuron circuits, have been carefully designed and optimized to realize key characteristics of a spiking neural network and a restricted Boltzmann machine such as asynchronous and parallel operation, spike-driven operation, and stochasticity. This work demonstrates the importance of co-optimization of device, circuit, architecture, and algorithm in neuromorphic processor design.

References

[1] https://openai.com/blog/ai-and-compute/

[2] G. W. Burr et al., "Neuromorphic computing using non-volatile memory," Adv. Phys. X, vol. 2, no. 1, pp. 89–124, Jan. 2017.

[3] M. Ishii, "On-Chip Trainable 1.4M 6T2R PCM Synaptic Array with 1.6K Stochastic LIF Neurons for Spiking RBM," in 2019 IEEE International Electron Devices Meeting (IEDM), 2019.

[4] C. Mead, "Neuromorphic electronic systems," Proc. IEEE, vol. 78, no. 10, pp. 1629–1636, 1990.

[5] T. Gokmen and Y. Vlasov, "Acceleration of Deep Neural Network Training with Resistive Cross-Point Devices: Design Considerations," Front. Neurosci., vol. 10, no. JUL, pp. 1–13, Jul. 2016.

[6] J. Woo and S. Yu, "Resistive Memory-Based Analog Synapse: The Pursuit for Linear and Symmetric Weight Update," IEEE Nanotechnol. Mag., vol. 12, no. July, pp. 36–44, 2018.

[7] H. Tsai, S. Ambrogio, P. Narayanan, R. M. Shelby, and G. W. Burr, "Recent progress in analog memory-based accelerators for deep learning," J. Phys. D. Appl. Phys., vol. 51, no. 28, p. 283001, Jul. 2018.

[8] Y. Kim, H. Kim, D. Ahn, J. Kim, "Input-Splitting of Large Neural Net- works for Power-Efficient Accelerator with Resistive Crossbar Memory Array," ISLPED, 2018.

[9] W. Maass, "Networks of spiking neurons: The third generation of neural network models," Neural Networks, vol. 10, no. 9, pp. 1659–1671, Dec. 1997.

[10] P. A. Merolla et al., "A million spiking-neuron integrated circuit with a scalable communication network and interface," Science, vol. 345, no. 6197, pp. 668–673, Aug. 2014.

[11] M. Davies et al., "Loihi: A Neuromorphic Manycore Processor with On-Chip Learning," IEEE Micro, vol. 38, no. 1, pp. 82–99, Jan. 2018.

[12] E. O. Neftci, "Stochastic neuromorphic learning machines for weakly labeled data," in 2016 IEEE 34th International Conference on Computer Design (ICCD), 2016, pp. 670–673.

[13] P. Cappelletti, "Non volatile memory evolution and revolution," in 2015 IEEE International Electron Devices Meeting (IEDM), 2015.

[14] A. Sebastian, M. Le Gallo, and E. Eleftheriou, "Computational phase-change memory: Beyond von Neumann computing," J. Phys. D. Appl. Phys., Aug. 2019.

SangBum Kim is an assistant professor at the Department of Materials Science and Engineering, Seoul National University. From 2010 to 2018, he was with the IBM T.J. Watson Research Center.

He is currently working on phase change memory devices for various memory applications such as storage-class memory, embedded memory, and brain-inspired neuromorphic computing.

He received a B.S. degree from Seoul National University, Seoul, Korea, in 2001 and a M.S. and Ph.D. degrees from Stanford University, Stanford, CA, in 2005 and 2010, respectively, all in electrical engineering. His Ph.D. dissertation focused on the scalability and reliability of phase change memory (PCM).