Pattern Training, Inference, and Regeneration Demonstration Using On-Chip Trainable Neuromorphic Chips for Spiking Restricted Boltzmann Machine

Uicheol Shin, Masatoshi Ishii, Atsuya Okazaki, Megumi Ito, Malte J. Rasch, Wanki Kim, Akiyo Nomura, Wonseok Choi, Dooyong Koh, Kohji Hosokawa, Matthew BrightSky,* Seiji Munetoh,* and SangBum Kim*

1. Introduction

The spiking neural network (SNN) has been proposed as the next generation of neural networks that could provide more computational power the deep learning neural networks widely use today.[1] However, studies are still searching for hardware implementation of SNN that can efficiently compute huge amounts of spatiotemporal spike trains.[2–4] To overcome the memory bottleneck of the von Neumann architecture, neuromorphic hardware has been devised by establishing an algorithmic neural network connection using analog nonvolatile memory (NVM) synaptic arrays.[5,6] Multiply-accumulate arithmetic, which is one of the core operations of the SNN, can be implemented in a fully parallel manner through a synaptic array of resistive devices for use in low-power and high-efficiency machine learning (ML) applications.[7–9] Beyond the multiply-accumulate calculation, in building the delicate peripheral neuron circuits that perform independent and asynchronous operations to process incoming spikes without delay, the application of the spike-timing-dependent plasticity (STDP) local learning rule on the synaptic array was presented.[10] Not limited to CMOS-based circuits, there also have been reported constructing

A fully silicon-integrated restricted Boltzmann machine (RBM) with an event-driven contrastive divergence (eCD) training algorithm is implemented using novel stochastic leaky integrate-and-fire (LIF) neuron circuits and six-transistor/2-PCM-resistor (6T2R) synaptic unit cells on 90 nm CMOS technology. To elaborate, designed a bidirectional, asynchronous, and parallel pulse-signaling scheme over an analog-weighted phase-change memory (PCM) synapse array to enable spike-timing-dependent plasticity (STDP) as a local weight update rule based on eCD is designed. Building upon the initial version of this work, significantly more experimental details are added, such as the on-chip characterization results of LIF and backward-LIF (BLIF) and stochasticity of our random walk circuitry. The experimental characterization of these on-chip stochastic neuron circuits shows a reasonable symmetry between LIF and BLIF as well as the necessary stochasticity for spiking RBM operation. Fully hardware-based image classification recorded 93% on-chip training accuracy from 100 handwritten MNIST digit images. In addition, we experimentally demonstrated the generative characteristics of the RBM by reconstructing partial patterns on hardware. As each synapse and neuron execute its computations in an asynchronous and fully parallel fashion, the chip can perform data-intensive machine learning (ML) tasks in a power-efficient manner and take advantage of the sparseness of spiking.
artificial neuron units[12] or dendrite networks[13] by combining resistive memristor technology for the implementation of elaborate STDP.[14,15] The SNN chip has been shown to lead to not only improvements in computing architecture by realizing fully parallel and asynchronous synaptic and neuronal operations but also low-power operations through sparse spiking activity on real hardware.[16] In this work, we implemented SNN hardware based on modified STDP with NVM arrays to operate low-power ML tasks suitable for edge computing devices.

In this work, we designed and fabricated an SNN chip capable of on-chip learning (Figure 1a,b) with a phase-change memory (PCM) synaptic array. The SNN chip operates as a restricted Boltzmann machine (RBM) with event-driven contrastive divergence (eCD).[17] An RBM is a two-layer bidirectional neural network with one visible layer and one hidden layer where the connectivity between neurons in the same layer is “restricted” (Figure 1c). Neftci et al.[17] presented the use of the eCD algorithm to train an RBM on SNN constructed with stochastic leaky integrate-and-fire (LIF) neurons using STDP weight update rule in an online and asynchronous fashion, which allows for simultaneous spiking activities from each LIF neuron on both layers. We implemented the algorithmic components of RBM and eCD by designing a unique hardware with an analog NVM-based synaptic cell array and peripheral neuron circuits (Figure 1d–g). The RBM has been proposed as a competitive ML algorithm that can be used for both supervised and unsupervised learning as well as in a wide variety of optimization problems.[18] Therefore, our low-power spiking RBM chip with densely integrated NVM synapses is also able to perform several ML applications in edge computing, where in situ real-time learning is currently challenging due to power and area limitations.

This article is a complementary work to our previous brief paper.[2] In the present work, we newly configured an field-programmable gate array (FPGA)-based demonstration system to flexibly operate the chip (Figure 1b, Xilinx Zynq-7000 SoC ZC706). The equipped FPGA efficiently transfers long spike trains to the SNN chip through on-chip scan chains. In this article, we added experimental data and discussions to experimentally characterize the working principles of the stochastic LIF circuitry functions in detail, as these are fundamental in the spiking RBM. We also demonstrated fully hardware-based MNIST image classification with the on-chip learning process, as well as pattern regeneration from the on-chip trained hardware spiking RBM network.

Figure 1. Spiking RBM neuromorphic hardware overview and components. a) Fabricated prototype chip micrograph. b) FPGA-based evaluation system. c) RBM network model and Poisson spike inputs. d) 6T2R PCM synapse cell circuit diagram with signal names. e) CAD layout image. f) TEM images. g) Array structure with 6T2R unit cells and LIF neurons circuits (2 × 2 is shown). h) 1000-states gradual SET with linearity-improved mushroom PCM cell with four examples of conductance evolution. i) median plots with +/− one standard derived by 50-time measurements of a PCM cell. Reproduced with permission.[2] Copyright 2022, IEEE Proceedings.
2. Experimental Section

2.1. 6T2R Unit Cell with Phase-Change Memory

PCM is one of the emerging memory technologies that involves storing data through resistance changes due to the phase transition of materials. Due to its advantages such as scalability and reliability, PCM technology has recently been commercialized with a sizable market. Their commercial production as a high-density storage class memory demonstrates the feasibility and maturity of PCM technology, which has led to studies exploring its use in diverse applications, such as automotive embedded memory, and neuromorphic computing. The use of PCM as synaptic devices in neuromorphic engineering has various attractive advantages such as nonvolatility, partial programmability, and a large on/off ratio. However, there is also a lack of linearity and symmetry as a synaptic device reduces the ML training accuracy in PCM-based neuromorphic devices. To overcome these issues, various compensation approaches have been proposed such as device, circuit, and algorithm optimization.

In this work, we implemented an RBM SNN chip with high-yielding PCM synaptic cells. Figure 1h,i shows that the PCM update linearity can be substantially improved by optimizing the programming conditions. The conductance of PCM can be gradually increased due to the accumulative properties implemented by partial SETs while the RESET process is abrupt. We adopted the strategy of using a pair of PCM cells to overcome the lack of linearity and symmetry of PCM caused by the abrupt RESET process.

We designed a novel 6T2R PCM unit cell to perform spiking RBM based on STDP, forward LIF (LIF), and backward LIF (BLIF), which are primitives of SNN (Figure 2a–c). Our previous 2T1R unit cell demonstrated fully asynchronous and massively parallel operations such as STDP and LIF in an area- and power-efficient manner. As the RBM is a bidirectional recurrent neural network, we added another transistor forming a 3T1R half unit cell for BLIF. We also implemented a signed synaptic weight by pairing two half unit cells with differential sensing. At the end of the bit line connecting each half unit cell, we further added a current mirror circuit to define the polarity of each half unit cell. Whenever the LIF_WL is activated, synaptic weight

![Figure 2](https://www.advancedsciencenews.com)

Figure 2. Circuitry operational design of the primitive functions for spiking RBM. Operation modes with simplified LIF neurons circuits; a) forward LIF (visible to hidden), b) backward LIF (hidden to visible), and c) STDP. Configurable pulse timings; d) visible neuron side, e) trimmable range for pulse delay and pulse width, and f) hidden neuron side. g) Configurable STDP_WL and its STDP function. Random walk function in LIF neuron circuits; h) random bit generation, i) random charge transfer circuits, j) example of random potential movement. k) A schematic diagram for bias spiking scheme. Bias neurons spike stochastically with their optimum average spiking rates \( N = 1 \) in this figure. Due to the exponential leakage of PSPs, bias spikes’ contribution to PSPs can be separated from that of others. As a result, bias spikes effectively increase the baseline of PSP. Reproduced with permission. © 2022 The Authors. Advanced Intelligent Systems published by Wiley-VCH GmbH
\[ w = G_{\text{total}} = G_p - G_m \] is provided for membrane potential integration.

After the LIF transistor is turned off, the STDP_WL pulse follows for weight update. When the hidden neurons fire due to enough LIF process, STDP_BL pulse occurred with time delay. These pulse schedules are designed for the implementation of the time correlation of neuron firing from pre- and postsynaptic neurons. Once the STDP_BL pulse delayed after neuron firing is temporally overlapped with STDP_WL pulse, the PCM cell is programmed with a large amount of current flowing through the PCM cell. The exact amount of current is determined by the STDP_WL pulse amplitude. As the pulse characteristics for unit PCM cell. The exact amount of current is determined by the STDP_WL pulse amplitude. As the pulse characteristics for unit cell operation, including STDP_WL_Gp, STDP_WL_Gm, and STDP_BL, are tunable (Figure 2d–f), we can precisely control the update timings of \( G_p \) and \( G_m \) by time correlatively performing SET or RESET programming. By combining them, we can implement modified STDP, a spike-based local weight update rule in RBM eCD (Figure 2g). During the data phase, by enabling external fire regardless of the membrane potential, specific spike trains are applied to visible neurons as input learning data. For the model phase, the external clamped data is turned off and the internal LIF serves to execute free running between the visible and hidden neurons. In each phase of eCD, the network state is optimized by updating through only the long-term potentiation (LTP) or long-term depression (LTD) of synaptic weights, respectively.\(^\text{17}^\) In the data phase, a small amplitude but long SET programming pulse is applied to STDP_WL_Gp, whereas a high voltage but short RESET programming pulse is entered into STDP_WL_Gm to implement positive weight update during internal LIF serves to execute free running between the visible and hidden neurons. When the hidden neurons

[2.2. Stochastic LIF Neuron Circuits]

The 6T2R unit cell is connected to a peripheral neuron circuit for LIF operation. Whenever LIF_WL pulse is generated from a visible neuron, the circuit connected with the PCM cell charge or discharge the membrane capacitor, which represents the membrane potential in the hidden neuron circuit. The firing probability of the stochastic neuron has a sigmoid-like varying tendency according to the membrane potential \[ \rho(u) = (\tau_r + \exp(-u))^{-1} \] where \( \rho, u, \) and \( \tau_r \) are the average firing probability, membrane potential, and refractory period, respectively.\(^\text{17,31}^\) This is a key feature of energy-based models such as RBM to migrate to a lower-energy state. To implement the stochasticity in the LIF neuron, we added a random walk function (Figure 2h–j), which is designed to implement the necessary neuronal behavior in the stochastic neural network, RBM.

The random walk is a system that has a probability distribution derived by several independent trials of randomly getting closer or farther when progressing to a certain state.\(^\text{132}^\) By applying this concept to the membrane potential of neuron circuitry, the membrane potential of each neuron circuit can be periodically (CCK<4>) increased or decreased by the fixed amount. The random bit (RAND_BIT), which determines whether to charge or discharge the membrane capacitor, can be assigned to 832 neurons and axons, respectively, using the two coupled LFSR and one XOR gate (Figure 2h). The XOR operations between two counterpropagating LFSRs generated from each prede fined seed provided RAND_BITS to each LIF neuron; this method could ensure temporally independent correlations between bit sequences.\(^\text{133}^\)

We operate the random walk circuits using pseudorandom bits generated by an on-chip linear feedback shift register (LFSR). The fixed amount of charges is then added (RAND_BIT=1) or subtracted (RAND_BIT=0) from the membrane potential capacitor using switched capacitors (Figure 2i,j). The 1.6 K stochastic LIF neuron circuits also perform neuronal operation functions such as the refractory period and the leaky function, as presented in previous works.\(^\text{2,211}^\)

[2.3. Trainable Biases]

In this spiking RBM chip, we implemented biases using bias neurons. The bias neurons exist on both layers and their functions are the same aside from the fact that they are always externally triggered to generate spikes with a given average spiking rate (bias spiking rate). Figure 2k illustrates how bias neurons implement a bias. Whenever bias neurons fire, the postsynaptic potential (PSP) of the corresponding postsynaptic neuron changes and the baseline PSP will change accordingly, as shown in Figure 2k. To optimize the training accuracy, biases need to be trained as well as weights on an extension of eCD.\(^\text{17}^\) As PCMs are updated based on the spike from adjacent neurons, the spiking rates of bias neurons affect the training of biases. Therefore, the optimal bias spiking rates that can be used to achieve maximum accuracy need to be determined.

As the influence of PCM conductance changes on PSPs should be consistent with that of their RBM counterparts, the average conductance change of PCM connected to a bias neuron and \( i \)-th visible (or hidden) neuron \( \Delta G_i \) for a small time interval \( \Delta t \) should be \[ \Delta G_i = \eta \sigma_{\text{STDP}} v_i \Delta t / N \sigma_{\text{ref}}^2. \] Here, \( N \) is the number of bias neurons; \( \eta \) is a learning rate; \( \sigma_{\text{STDP}} \) is a learning window; \( \tau_{\text{ref}} \) is a refractory period; and \( v_i \) are the average spiking rates of the bias neurons and the \( i \)-th visible (or hidden) neuron, respectively. Also, from learning rule of 6T2R spiking RBM, the average conductance change is \[ \Delta G_i = \eta \sigma_{\text{STDP}} v_i \Delta t. \] Thus, by combining those two equations, the optimum average spiking rate of bias neurons could be found to be \[ \alpha = N^{-1/2} \sigma_{\text{ref}}^{-1}. \] Altogether, trainable biases were implemented in terms of bias neurons, and bias training was also properly enabled by setting the average spiking rates to their optimum values.

[3. Results and Discussion]

[3.1. LIF and BLIF]

We tested the LIF operation by accessing a unit cell on a LIF_WL with 832 synaptic unit cells. By programming PCM devices on the selected unit cell, we were able to investigate the impact of the total conductance \( G_{\text{total}} \) on LIF. We measured LIF spike outputs
by applying 200 external input spikes through the visible neuron circuit. We reported the change in LIF outputs according to $G_p$ and $G_m$. As shown in Figure 3a,b, the number of output spikes increases as the positive conductance, $G_p$, increases through repeated SET programming. When the conductance of the $G_p$ PCM cell increases, the discharge amount of the membrane capacitor in the LIF circuit is increased as well, so it is possible to fire with an even smaller number of input spikes, thus resulting in more output spikes. On the other hand, increasing the conductance of the $G_m$ PCM, which represents negative conductance, decreases the $G_{total}$ of the unit cell. This indicates that more incoming spikes are required for the membrane potential to exceed the predetermined threshold voltage, thus resulting in a decrease in the number of output spikes. We measured the LIF output by adjusting the interval of input spikes, which demonstrates that the leak function and refractory time are properly working in our LIF circuitry. The larger spike interval generates less LIF output due to decay of the membrane potential implemented by the leak function, whereas if the interval is too small, the output is measured as small because many input spikes are ignored by the refractory time after firing.

We also proceeded with identical experiments on the hidden-side neuron circuit to compare LIF and BLIF, and we confirmed that bidirectional connections were implemented in the fabricated circuitry (Figure 3c.d). In addition to the hardware in situ results plotted with dot symbols, we performed a numerical simulation (the results are plotted with lines) of the ideal LIF-BLIF circuitry system with the refractory period. The well-fitted results in Figure 3c,d show a comparable tendency of output spikes, thus indicating that our LIF circuitry is well-fabricated. By demonstrating the symmetrical performance of LIF and BLIF, the neuromorphic hardware can implement spiking RBM that requires symmetry between forward and backward propagation.

3.2. Random Walk

We used the random walk circuit to implement the stochastic neuronal behavior required by RBM. For the hardware in situ experiment, we counted the number of output spikes while operating random walk circuits for 3,328 clocks. We derived the firing rate by measuring the number of output spikes that are generated. As illustrated in Figure 4a, the sigmoid-like firing probability is determined by the distance between the threshold voltage and the initial membrane potential, thus indicating that the neuron has nonzero firing probability even without any incoming spikes. Further, the amount of change in the membrane potential can be fine tuned by adjusting the voltage levels of VP_DN and VN_UP. The random walk step size eventually controls the stochasticity of neuron behavior; optimizing this behavior is one way to maximize the performance in the RBM algorithm.

3.3. MNIST On-Chip Training

Figure 1c shows an implemented RBM model on the fabricated SNN chip and a digit classification scheme. In the training process, binarized MNIST database images\(^{[34]}\) are converted into Poisson spike trains by C++-based code with applying predefined spike rate and refractory period and transmitted to the network as external inputs via the equipped FPGA system. During the data phase, 524 visible neurons, 30 sets of 10 label neurons, and 8 bias neurons are triggered by the external inputs with average spike rate of 200 Hz. This is followed by the model phase, during which the external inputs are turned off except for bias neurons, thus generating spikes only by the internal LIF and BLIF between visible and hidden neurons. For the inference, Poisson spike trains are transmitted only to the data neurons excluding label neurons. The well-trained RBM internally triggers specific label neurons to fire that correspond to the shown MNIST image. We can detect which label neuron has the most spikes using on-chip counter circuits and determine the accuracy of inference.

Figure 4b illustrates the experimental result of the MNIST on-chip training. The 100-image inference was performed with 100 trained images after the training of 100 images per epoch, resulting in 93.0% maximum accuracy. The 10 000-image test is trained with 2,000 randomly chosen images per epoch from 60 000 training images, and it scored the best accuracy of 73.57%. The degradation of accuracy with an increasing number of test images can be explained by the inherent issues with the analog neuron circuits and synaptic devices. PCM as a synaptic device exhibits an abrupt conductance change in the RESET
process, which is expected to cause undesirable influences on weight update. To mitigate this, we have reduced the pulse width of STDP_WL_Gm in the data phase and STDP_WL_Gp in the model phase so that RESET occurs less frequently than SET. In other words, through the pulse width control, the adverse impact of the abrupt RESET, can be alleviated by reducing the RESET probability. In addition, the use of the previously presented "PCM Refresh method" can improve the accuracy even further. These algorithmic solutions indicate that our silicon-integrated neuromorphic hardware has the potential for further accuracy improvements. Power consumption was estimated by SPICE simulation while considering 90 nm fabricated design and 20 Hz average spike rate, ultimately deriving 8.95 pJ per synaptic spike operation with active and static power of 4.10 and 4.85 pJ, respectively. To our knowledge (Table 1), this is the first SNN demonstration of on-chip training with fully silicon-integrated neuron circuits and a large-scale (more than 1 million) PCM synaptic devices implementing a well-established and practical ML algorithm (RBM).

3.4. Pattern Regeneration

The well-trained RBM network can reconstruct an originally trained pattern based on an incomplete input pattern. With this generative capability, spike trains of label neurons are reconstructed by receiving the Poisson spike trains of the input image in the inference step. Similarly, for the regeneration step, entering the Poisson spike trains of an incomplete image with

Table 1. Comparison with recent works.

| Synaptic cell | On-chip training | Spiking or Nonspiking | Demonstration | Number of synapses |
|---------------|------------------|----------------------|---------------|--------------------|
| This work     | PCM              | Yes (in situ)        | Spiking       | Fully silicon integrated | 692 K/core |
| Mochida et al.[37] | ReRAM          | No                   | Nonspiking    | Fully silicon integrated | 4M         |
| Ambrogio et al.[38] | PCM            | Yes (simulation)    | Spiking       | Simulation only     | N/A        |
| Ambrogio et al.[39] | PCM + 3T1C     | Yes (H/W, S/W mixed) | Nonspiking    | H/W, S/W mixed      | 205 K      |
| Merolla et al.[16] | SRAM           | No                   | Spiking       | Fully silicon integrated | 64 K/core, 256M/chip |
true label results in a complete image (Figure 4c). By screening one half of an MNIST image, only 236 neurons among 524 data neurons are externally triggered, and the remaining data neurons are used to read reconstructed spike trains. The demonstration was performed after 100-image on-chip training with 90% accuracy. Figure 4d shows the reconstructed MNIST images. All images are from 30 different original images. From this experiment, we confirmed that the hardware has successfully implemented spiking RBM by demonstrating not only the capability of pattern recognition but also the performance of pattern regeneration based on the generative artificial neural network.

4. Conclusion

In this work, we designed and demonstrated an SNN processor at 90-nm node integrating a novel 6-transistor/2-resistor (6T2R) PCM synaptic cell arrays storing analog weights and asynchronous stochastic neurons with on-chip training and inference capabilities to perform data-intensive ML tasks. We experimentally demonstrated that the core algorithmic functions of spiking RBM can be implemented by primitives such as LIF, BLIF, STDP weight update, random walk, and the refractory timing system. Experimental handwritten image classification using 100 samples from the MNIST database shows 93% training accuracy. With the on-chip-trained RBM network, we also demonstrated pattern reconstruction.

This experiment demonstration lays the groundwork for a power-efficient SNN processor with an array of analog synaptic devices that are capable of fully asynchronous and parallel operations that facilitate the implementation of wide and deep SNNs. The performance and power consumption of the proposed SNN processor can be further improved through the development of more innovative stochastic neuron devices as well as the improvement of synaptic devices in weight-update symmetry and linearity, along with power and area efficiency. In the near future, we can expect better on-chip learning performance through improvements in the analog characteristics at the device-level[27] as well as the application of the “PCM refresh method” at the algorithmic system-level.

Acknowledgements

U.S. and M.I. contributed equally to this work. The authors would like to thank the IBM PCM group and the microelectronics research lab (MRL) for manufacturing support, as well as Maxence Bouvier for phase-change memory measurement support. This research was conducted under IBM-SNU Joint Study Agreement. Part of this work conducted by the SNU researchers is supported by the New Faculty Startup Fund from Seoul National University, the LG Display (LGD-SNU Incubation Program), and National R&D Program through the National Research Foundation of Korea (NRF) funded by Ministry of Science and ICT (NRF-2020R1F1A1076565). This article was amended on May 24, 2022 to annotate M.B. and S.M. as co-corresponding authors.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

Research data are not shared.

Keywords

image classifications, neuromorphic computing, pattern regeneration, phase-change memory, spiking neural networks, restricted Boltzmann machines, stochastic leaky-integrate-and-fire

Received: February 3, 2022
Revised: April 18, 2022
Published online: May 22, 2022
