

A Memristor based Unsupervised Neuromorphic System Towards Fast and Energy-Efficient GAN

Fuqiang Liu and Chenchen Liu
fqliu92@gmail.com chchliu@clarkson.edu

ABSTRACT

Deep Learning has gained immense success in pushing today's artificial intelligence forward. To solve the challenge of limited labeled data in the supervised learning world, unsupervised learning has been proposed years ago while low accuracy hinders its realistic applications. Generative adversarial network (GAN) emerges as an unsupervised learning approach with promising accuracy and are under extensively study. However, the execution of GAN is extremely memory and computation intensive and results in ultra-low speed and high-power consumption. In this work, we proposed a holistic solution for fast and energy-efficient GAN computation through a memristor-based neuromorphic system. First, we exploited a hardware and software co-design approach to map the computation blocks in GAN efficiently. We also proposed an efficient data flow for optimal parallelism training and testing, depending on the computation correlations between different computing blocks. To compute the unique and complex loss of GAN, we developed a diff-block with optimized accuracy and performance. The experiment results on big data show that our design achieves $2.8\times$ speedup and $6.1\times$ energy-saving compared with the traditional GPU accelerator, as well as $5.5\times$ speedup and $1.4\times$ energy-saving compared with the previous FPGA-based accelerator.

1. INTRODUCTION

Deep learning has achieved great success in various artificial intelligence applications such as object detection [1, 2], tracking [3] and natural language processing [4]. Normally, supervised learning is employed in the state-of-the-art applications, where a deep neural network is trained from labeled training data and desired outputs are obtained after going through an inferring function like backpropagation [5, 6]. The supervised learning has been proved powerful, however, challenges emerged with the fast growth of the applications complexity: large-scale of labeled data is in demand and its learning capability is constrained.

Unsupervised learning, which has the capability to learn from the unlabeled data directly appears as a possible solution. Nonetheless, accuracy is usually low in the conventional unsupervised learning and its feasibility in the realistic applications is hindered [7]. Recently, generative adversarial networks (GAN) was proposed as a promising solution for the above challenges [8]. GAN can estimate generative models from an adversarial process by training two models—generative and discriminate model simultaneously [8]. Feature representations can be learned from unlabeled data and improved accuracy of unsupervised learning is achieved on GAN [9]. However, the challenge of high demand for computing resource that exists in deep neural network becomes more severe in GAN computation. Both training and testing executions are particularly slow and energy hungry [8, 9].

Extensively research efforts have been devoted to accel-

ate and improve computing efficiency of deep learning, such GPU [6], FPGA [10], emerging non-von Neumann accelerators [11, 12]. Very recently, GPU [9] and FPGA [13] based GAN computations have been proposed with significantly improved speed and energy efficiency compared with CPU. However, the computing efficiency is still constrained as data bandwidth in these architectures is still limited and the performance improvement mainly relies on reckless resource accumulation. Recent studies in developing novel non-von Neumann accelerators such as in-memory accelerator [14, 15] and neuromorphic computing system [16, 12, 17], esp. the designs with novel nano-devices like memristor paved a way towards computational efficient GAN.

In [11, 12], convolutional neural networks (CNNs) were deployed on memristor crossbar and an on-line training process with backpropagation were implemented via a hardware and software co-design methodology. However, these previous approaches cannot be used directly in GAN computation because of the following reasons. First, different with traditional CNNs training, two learning models are executed simultaneously in GAN's training phase and requires an adaptive data flow for optimal computing efficiency. Moreover, backpropagation is simplified in a hardware-friendly way that the initial error in training is considered as the difference between the true label and the predicated label, which works conditionally under the prediction that the cost function of CNN is cross entropy. This assumption is invalid in GAN, therefore, the memristor-based hardware implementation in previous works cannot be utilized.

In this work, we developed a memristor-based unsupervised neuromorphic system for a fast and energy-efficient GAN computation to solve the above challenges. Our contributions can be summarized as follows:

- We exploited a hardware and software co-design approach to map the computation blocks in GAN to the memristor-based crossbars efficiently. The computing system is composed of three major blocks—*Generator*, *Discriminator*, and *Diff* block. The *Diff* block is designed to compute the cost function of GAN accurately with low hardware cost;
- We proposed an adaptive data flow for GAN with optimal computation parallelism and efficiency. In forward phase, the *Generator* and *Discriminator* block worked in parallel to generate artificial data and extract features; In backward phase, the *Generator* and *Discriminator* block were trained effectively with the initial errors computed by the Diff block.
- We evaluated the system accuracy in different data precision and the system performance in speed and energy. The proposed system performance on ImageNet and Lsun/bedroom was also compared with the GPU-based and FPGA-based GAN computation in previous works.

The experimental results show that our proposed design can achieve $2.8\times$ ($2.7\times$) and $5.5\times$ ($4.8\times$) speedup, as well as $6.1\times$ ($6.1\times$) and $1.4\times$ ($1.5\times$) energy-saving compared with GPU-based [9] and FPGA-based [13] GAN accelerators respectively on Lsun (ImageNet) dataset.

2. BACKGROUND

2.1 Generative Adversarial Networks

The generative adversarial network was developed as an unsupervised model that can learn effective feature representations from unlabeled data while having improved accuracy compared with traditional unsupervised learning [8, 9]. Two learning models form the GAN: a generator and a discriminator. Normally, the generator is a deconvolutional neural network (Deconv-NN) for artificial data generation, and the discriminator is a convolutional neural network (CNN) for distinguishing the artificial data from real data. The training phase of GAN involves two major learning procedure—effective generator and discriminator learning based on back-propagation. The target of the training process is obtaining a generator that can generate most likely the same as the true CNN training data and a discriminator that can extract feature effectively. The training process is can be summarized as four major procedures, which are named as $D_{forward}$, D_{back} , $G_{forward}$, and G_{back} in this work.

- $D_{forward}$ computes the cost function to obtain the error that should be transmitted to the discriminator for its backward weight updating. More specifically, with a batch of m noise samples— z^1, \dots, z^m are given as inputs, the generator generates m artificial samples and this process is defined as $G(z^i)$. The CNN based discriminator processes the m artificial samples (i.e. $G(z^i)$) and m real samples (i.e. x^1, \dots, x^m) through forward computations, and then cost function is computed following Equation 1.

$$Error_D = \nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^m [\log D(x^i) + \log(1 - D(G(z^i)))] \quad (1)$$

- D_{back} updates the weights of the discriminator by ascending the stochastic gradient obtained from Equation 1, i.e. $Error_D$.
- $G_{forward}$ computes the cost function to gain the error that should be given to the generator for its weight updating. The cost function is computed as Equation 2.

$$Error_G = \nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^m \log(1 - D(G(z^i))) \quad (2)$$

- G_{back} updates the weights of the generator by ascending the its stochastic gradient obtained from Equation 2, i.e. $Error_G$.

Here, $G(\cdot)$ and $D(\cdot)$ represents the generator and discriminator respectively, x is the real data, and z is the noise given to the generator.

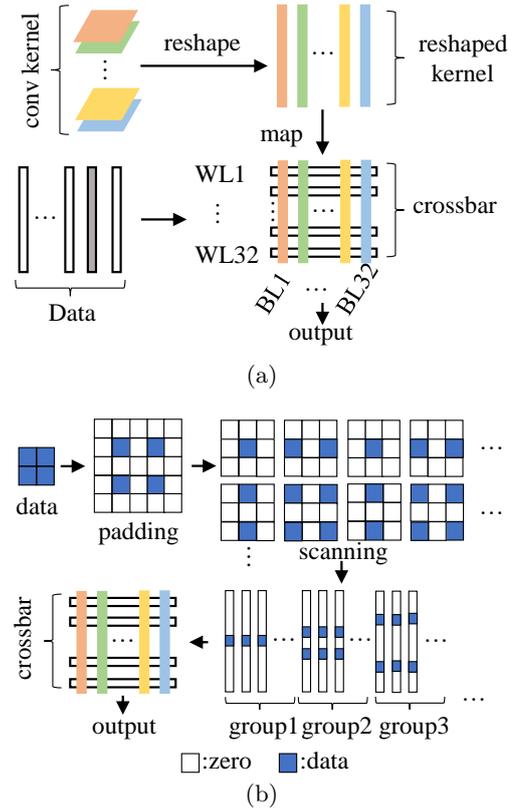


Figure 1: Implementation of CNN (a) and Deconv-NN (b) on memristor crossbar

2.2 Memristor Crossbar for (De)Convolutional Computation

The limited data bandwidth, as well as the performance gap between processing units and memory of the conventional computing platform becomes a major obstacle in the deep learning based applications. Novel computing platforms, such as the in-memory accelerator [14], neuromorphic computing [18], etc. therefore have been extensively investigated as a promising solution. The emerging of novel nano-devices such as spin, phase change and memristor device also accelerates its development and corresponding accelerators are developed accordingly. Among them, the memristor based computing platform attracts people's attention own to the high density, high speed, multiple level states, etc [19, 11, 12]. In this work, the memristor-based computation platform for GAN is developed. The generator that based on deconvolutional network and the discriminator based on convolution network of GAN is deployed on the memristor crossbar structure.

In previous work, convolutional networks have been deployed on the memristor-based crossbar structure, as is depicted in Figure 1 (a) [12]. For example, to deploy a convolutional layer with 32 kernels, each kernel is reshaped to a vector that can be programmed to a memristor crossbar. And the inputs data is given to the memristor crossbar to execute the dot-products computations. Multiple crossbars are connected in parallel to form the large-scale convolutional layer because of the size limitation of the memristor crossbar [12]. The ReLU activation function of CNN can be

implemented by the integrate-and-fire circuits (IFCs) and other digital logic [16, 12]. The deployment of deconvolutional network is similar to the convolutional layer. One major difference is the data should be zero-padded before giving to the crossbar, as is shown in Figure 1 (b). To optimally decrease the executions involved by the zeros, we transform and group the input vectors with the zeros in the same locations, and only non-zero inputs in the rows are given to the crossbar and computed.

3. DESIGN METHODOLOGY

In this section, we built a memristor-based neuromorphic design for accelerating both training and testing phase of GAN. The basic computing architecture and data flow is described in subsection 3.1 To achieve optimal system performance, we proposed a cross-parallel pipeline execution flow which is introduced in 3.2.

3.1 Memristor-based GAN Architecture

Based on the GAN computation described in 2.1, we developed the memristor-based computing architecture of GAN computation. As is shown in Figure 2, the proposed architecture is composed by four integrated components: Generator block, Discriminator block, Diff block, and Control unit. The functionality of these components can be summarized as below.

(1) The Discriminator block is designed to compute the $D(x^i)$ and $D(G(z^i))$ in Equation 1 for the stochastic gradient calculation. It is composed of a sea of connected memristor-based CNN units;

(2) The Generator block is built to calculate the $G(z^i)$ in generating the artificial samples from the noise inputs for computing the stochastic gradient in generator weight updating. Similarly, it is composed of a sea of connected memristor-based DeCNN units;

(3) Following Equation 1 and Equation 2, the Diff block computes the gradients of the discriminator and generator block respectively. The Diff block is constructed by the memristor-based circuits blocks including LUT (look up table), memory, adder, etc.

(4) The Control unit is designed to control the data flow and is built with combinational and sequential digital logic.

In Figure 2, the basic data flow is demonstrated as $a \sim f$. The explanation of each step is summarized as below, where

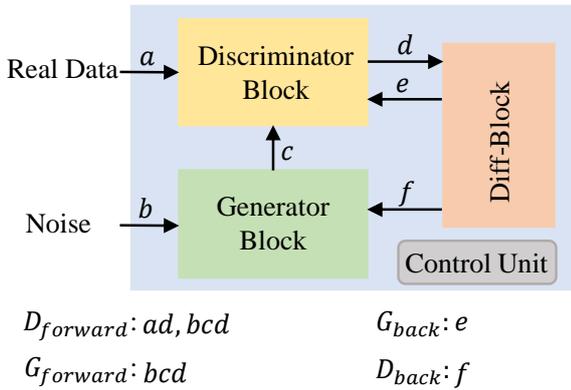


Figure 2: Architecture of the proposed memristor-based GAN computation.

the definitions of $D(\cdot)$, $G(\cdot)$, x , z , and m are as the same as those in subsection 2.1.

- a : The real data with m samples in a batch, i.e. x^i is given to the Discriminator block to compute the $D(x^i)$, where $i \in \{1, \dots, m\}$.
- b : The noise data with m samples in a batch, i.e. z^i is given to the Generator block to obtain artificial samples $G(z^i)$, where $i \in \{1, \dots, m\}$.
- c : The $G(z^i)$ is generated from the Generator block and then transmitted to the Discriminator block to compute the $D(G(z^i))$, where $i \in \{1, \dots, m\}$.
- d : This step computes the transmits the $D(x^i)$ and $D(G(z^i))$ from the Discriminator block to the Diff block. The above steps $a \sim d$ fulfill the forward computations of GAN.
- e : Based on the gradient calculated by the Diff block, i.e. $Error_D$ in Equation 1, this step updates the weights of the Discriminator.
- f : The weights of the generator is updated in this step according to the gradient obtained from the Diff block. These two steps e and f implement the backward computations in GAN.

Correspondingly, the $D_{forward}$ is composed of a, c and b, c, d , D_{back} is fulfilled by e , $G_{forward}$ is implemented by b, c, d , and G_{back} is the data flow of f .

3.2 The Cross-Parallel Pipeline

In this section, we proposed a cross-parallel pipeline based on the basic data flow for optimized computing efficiency. Hence, we first analyze the time cost of each step among $a \sim f$ in Section 3.1. The analysis is executed based on the simulations on NVSim simulator [20] and the results are demonstrated in Figure 3. In the analysis, the steps of e and f are divide into two categories: data transmission and data computation. For example, e_1 refers to the gradients transmission time from the Diff block to the Discriminator block while e_2 represents the time consumed by updating the weights of the discriminator. Similarly, the f_1 is the gradients transmission and f_2 is the cost of weights updating

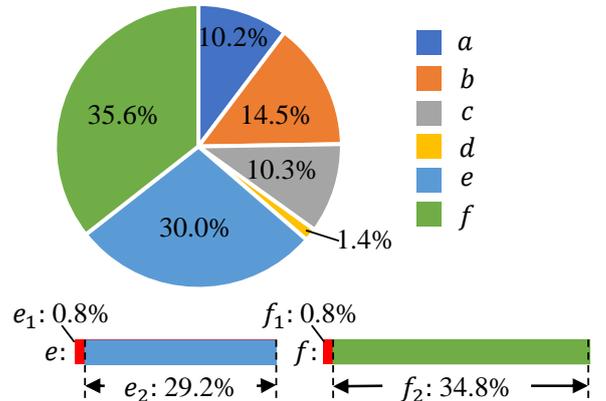


Figure 3: Time cost of each step of the basic data flow.

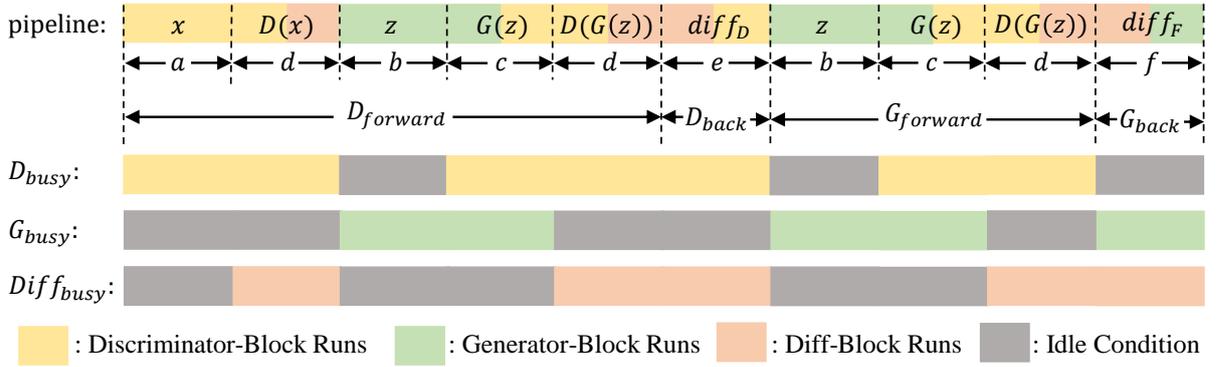


Figure 4: The pipeline of each computing block based on the basic data flow.

of the generator. It is observed that the majority of the time cost occurs in updating the generator and discriminator.

Figure 4 shows the pipeline of each computing block, i.e. the Discriminator block, Generator block, and Diff block based on the basic data flow. As is described, different colors represent the execution states of each block. Then, we analyze the independence property each execution in these blocks to build the cross-parallel pipeline, where the independent executions are designed to compute in parallel. For example, observed from Figure 4, the discriminator block and the generator block are independent from each other when the a and e_2 are executing by the discriminator block or b and f_2 are executing by the generator block. Hence, these computations can be optimized to be parallelism data flow. There are only two conditions that the generator and discriminator block cannot work in parallel: first, during b and c executions; second, the weights updating has not finished. Otherwise, the generator and discriminator can work highly in parallel and the optimized cross-parallel pipeline is depicted in Figure 5.

In the developed cross-parallel pipeline, the $D_{forward}$ is divided into two parts— $D_{forward}^{t,1}$ and $D_{forward}^{t,2}$, where t represents the t^{th} iteration in GAN training. $D_{forward}^{t,1}$ includes the basic a and d execution and the $D_{forward}^{t,2}$ refers to the basic $b \sim d$ processing in the discriminator block. In the cross-parallel pipeline, the a (i.e. real samples processing) and the b (i.e. artificial samples generation) are exe-

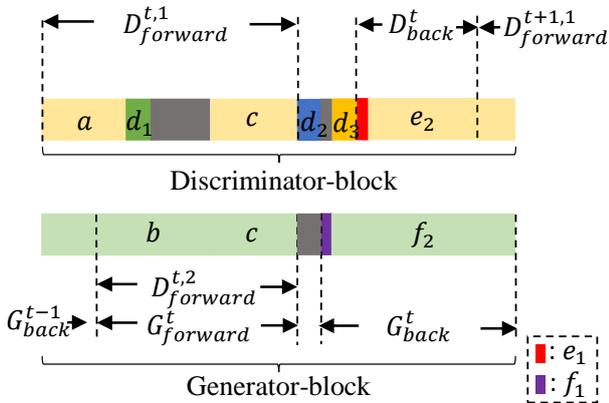


Figure 5: The cross-parallel pipeline.

Table 1: Computation time of Basic Pipeline vs. Cross-Parallel Pipeline

	Basic Pipeline	Cross-Parallel
Each Iteration(s)	0.18	0.11
$D_{IdleTime}(s)$	0.069	0.018
$G_{IdleTime}(s)$	0.033	0.015

cuted simultaneously. It is observed from Figure 3 that b is more time-consuming than a , hence the $D(x^i)$ that computed from a is transmitted and stored in the Diff block firstly which is represented by d_1 in Figure 5. And a memory unit based on memristor crossbar is designed in the Diff block to store the computing results of $D(x^i)$. Then, the discriminator is idle which means no execution occurs until the artificial samples generation (i.e. b) is finished. The $D(G(z))$ is transmitted to Diff block after the c finished, which is represented by d_2 in Figure 5. Immediately, d_3 which includes the gradient computation and its transmission to the discriminator block is executed. Consequently, the weights updating in the discriminator (e_2) and generator (f_2) block run simultaneously. In addition, although the e_2 is executed faster than f_2 , the next training iteration of the discriminator starts asynchronously without introducing additional memory usage in the discriminator block as is indicated by $D_{forward}^{t+1,1}$. The training process of the generator and discriminator become synchronous before c starts in the $(t+1)^{th}$ iteration.

As is discussed above, the parallelism of the computation can be largely improved. The time cost of the GAN computation following the basic pipeline and the proposed cross-parallel pipeline is evaluated on CIFAR-10 dataset [21] to indicate this improvement. Also, the simulation is executed on the simulator NVSim [20]. The results are shown in Table 1, we can observe that $1.6\times$ speedup is achieved by utilizing the cross-parallel pipeline. In addition, the usage rates of the Discriminator block and Generator block are improved $3.8\times$ and $2.2\times$ respectively.

4. DESIGN DETAIL

In this section, the implementation of the Discriminator block, Generator block and Diff block is explained in detail. All of the above blocks are implemented based on the memristor crossbars.

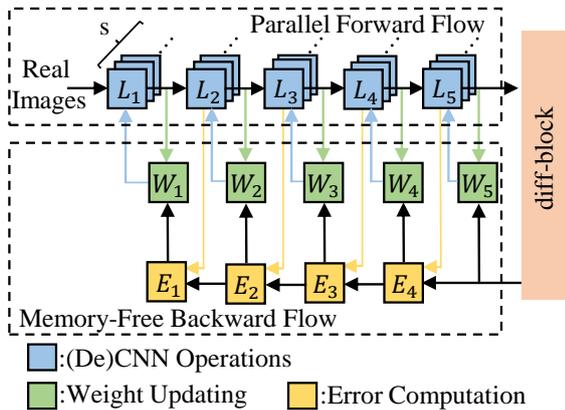


Figure 6: The parallel and memory-free structure for the discriminator and generator blocks. The black arrows represent data input and output, while the colorful arrows represent the programming.

4.1 Parallel and Memory-Free Structure for Discriminator and Generator Blocks

In previous research [12], the memristor-based computations for CNN training and testing has been implemented. However, these designs cannot be used in the GAN training because two major reasons. First, the gradient for the output layer of CNN was computed as the output minus the true label previously, which is not adapted to GAN. Moreover, the previous framework involves a large number of memory units, such a design will result in heavy area cost and time-consuming in GAN computations.

To solve the above challenges, we proposed a parallel and memory-free structure as is shown in Figure 6. The squares in different color represent the processing units of error computations, weight updating, and (de)convolutional operations respectively. These units are built by the memristor-based crossbar, IFC, and digital logic [16, 12], and a simplified mapping scheme is shown in Figure 7. The proposed structure is composed of a parallel forward flow and a memory-free backward flow. It can work as the discriminator or generator block with different initial weights programmed on the (de)convolutional operation units.

4.1.1 The Parallel Forward Flow

The forward processing of GAN includes the CNN and DeCNN computations. Consider a GAN structure with 5-layer CNN and 5-layer DeCNN, the parallel flow is depicted in the above part of Figure 6. Initial or updated weights, i.e. w_l in Figure 7 (a) are programmed to the memristor crossbar of the (de)convolutional layer and the results processed from the forward flow are obtained as o_l that works as the input to the next layer. The mapping and programming method is follows Section 2.2 and previous researches [12, 16]. Multiple samples can be processed on these same processing units with different inputs, and thus parallel computations can be achieved. The final output of the parallel forward flow is transmitted to the Diff block.

4.1.2 The Memory-Free Backward Flow

The memory-free backward flow aims to update the weights of the forward flow based on the gradient from the diff-block.

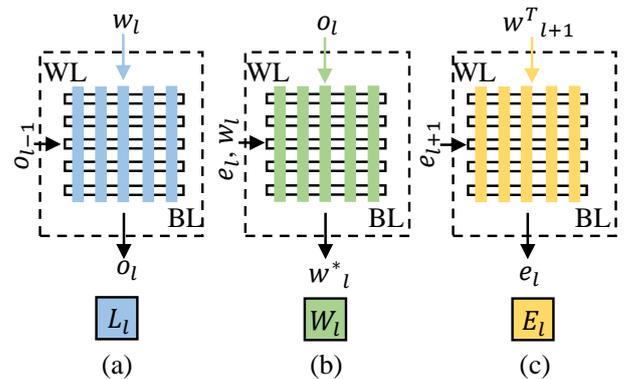


Figure 7: The simplified mapping method of different layers. Blue: (de)convolutional operations, Green: weight updating, Yellow: error computation.

The backpropagation method used to update weights of (de)cnn can be summarized as equation 3, where l represents the index of the (de)convolutional layer, e_l represents the gradient, w_l represents the weights, w_l^* represents the updated weights and w_l^T represents the transposition matrix of the weights, o_l' represents the derivative of the l^{th} layer output signals and α represents the learning rate. Because the derivative of ReLU activation function is equal to the output signal itself, o_l' is equal to o_l .

$$\begin{cases} e_l = e_{l+1} \otimes w_l^T & (a) \\ w_l^* = \alpha \cdot e_l \otimes o_l' + w_l & (b) \end{cases} \quad (3)$$

The memory-free backward flow is proposed to implement the backpropagation method, which is depicted as the below part of figure 6. Weights' transposition, i.e. w_l^T in Figure 7 (c) are read from the operation units in the forward flow and programmed to the memristor crossbar of the error computation units. Outputs of each layer in the parallel forward flow, i.e. o_l in Figure 7 (b) are programmed to the memristor crossbar of the weight updating units. The input to the error computation unit is the errors computed from the error computation unit in the next layer, i.e. e_{l+1} in Figure 7 (c) and then current layers' errors, i.e. e_l in Figure 7 (b) are computed. These errors are also input to the weight updating units to compute the updated weights i.e. w_l^* in Figure 7 (b).

In previous designs [12, 16], they use memory to store the updated weights or the inter-layer signals of CNN. The reason that the proposed design does not need special memory is that we store the weights and inter-layer signals in the same memristor crossbar as computation units. The inter-layer signals are programmed to the weight updating units directly used to compute the updated weights, shown as the green arrows in figure 6. The updated weights are programmed to the (de)cnn computation units, shown as the blue arrows in figure 6.

4.1.3 Timing Sequence of the Parallel and Memory-Free Structure

The timing sequence of one iteration for the parallel and memory-free structure is detailed as Figure 8. First, (de)cnn operation units process the input data and the inter-layer

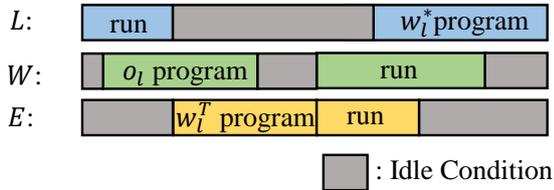


Figure 8: The timing sequence of the parallel and memory-free structure

signal is programmed to the weight updating units immediately. After (de)cnn operation units processing the input data, the weights from these units are programmed to the error computation units and then they begin computing the backpropagation errors and weight updating units begin computing the new weights which are programmed to the (de)cnn operation units directly.

4.2 The Implementation of Diff-Block

The implementation of Diff-Block is detailed shown as Figure 9. Diff-block is composed by two memristor-based look-up tables (LUT) [22], the memristor-based memory unit [20, 23] and two memristor-based adders [16]. LUT1 stores values of $\nabla_{\theta} \log D(x)$ and LUT2 stores values of $\nabla_{\theta} \log(1 - D(G(z)))$. Linear transformation, $\frac{1}{m}$, can be done in the adder [16].

When the discriminator-block transmitted $D(x)$ to the diff-block, values of $\nabla_{\theta} \log D(x)$ are read from the LUT1 and stored in the memristor-based memory. The memory should be able to store m values where m represents the batch size. Generally, the batch size is 64. When the discriminator-block transmitted $D(G(z))$ to the diff-block, values of $\nabla_{\theta} \log(1 - D(G(z)))$ are read from LUT2. Those values are input to the adder and the $Error_G$ is computed. Meantime, the data in the memory are input to the adder as well as the above result and $Error_D$ is computed.

5. EVALUATION

In this section, we evaluate the performance of the proposed memristor-based GAN accelerator on accuracy, speed, energy, and area cost. The performance is compared with the previous GPU-based platform [9] and the microarchtural design [13] based on FPGA. The Nvidia Geforce GTX 1080 is used as the GPU platform. The proposed design is evaluated on NVSim [20] simulator environment. The memristor crossbar size is designed to be 32×32 , the resistance range of the memristor device is set to be $[50K\Omega, 1M\Omega]$, and the required crossbar number is calculated following the implementation in 2.2 and [11]. The circuits designs for the neurons and control units follows the [12].

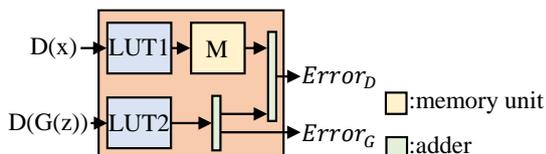


Figure 9: The architecture of diff-block

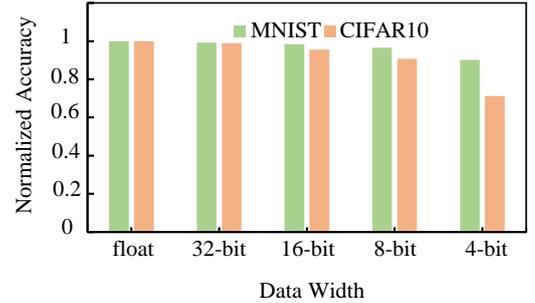


Figure 10: Accuracy in different data precision.

5.1 Accuracy in Different Data Precision

In general, the memristor supports limited precision in data transmission, data storage and computation [24, 16]. Although analog states of memristor have been reported by the HP Research Lab [25], high precision involves in scariification of speed and design cost. In this section, the data bitwidth for optimized accuracy and design cost is explored, and the GAN computing accuracy in different data bitwidth is shown in Figure 10. As referred in [9], the performance of GAN can be measured by using the discriminator as a feature extractor for a classifier. In this experiment, SVM works as the classifier and the discriminator of GAN works as the extractor for image classification. MNIST and Cifar-10, whose characters are listed as table 2, are used as the benchmark. The accuracy based on the feature extractor trained from GPU is regarded as the baseline. The accuracy based on the fixed-point feature extractors is compared with the baseline. Two dataset is utilized as is depicted in Table 2: MNIST and CIFAR-10.

Figure 10 shows the normalized accuracy when discriminators are trained by data in different data format and bitwidth. The results show that the system accuracy with 8-bit data precision has a slight accuracy loss the the normalized accuracy is still higher than 90% on MNIST and CIFAR10. However, significant accuracy loss is introduced in the 4-bit data precision. Therefore, the 8-bit data precision, i.e. memristor device with 8-bit states is utilized in this work and the following evaluation.

5.2 Computing Parallelism

In the training process of GAN, the generator processes a batch of noise samples and the discriminator processes a batch of real images. The computing parallelism is referred to the generator (or discriminator) blocks that compute in parallel in each training iteration, i.e. the s in Figure 6 As is indicated in Section 3.2, the computing efficiency can be improved by higher computing parallelism. Note that higher parallelism results in an increase of the area and design cost in the proposed GAN computing system. In this section, the speed and area cost in different data parallelism scenarios

Table 2: Dataset

	training size	testing size	scale	category
MNIST	60000	10000	28×28	10
CIFAR-10	50000	10000	32×32	10

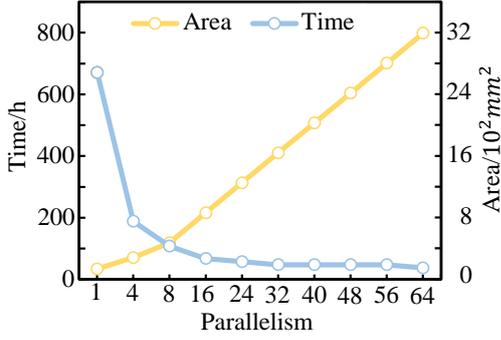


Figure 11: Time and area cost on different computing parallelism

are explored. The bedroom dataset from Lsun [26] is used in this evaluation. The maximum parallelism size is selected to be 64.

Figure 11 shows the time and area cost in different parallelism. The computation time decreases with the increase of the parallelism size while the area cost increases heavily in high parallelism scenario. The main reason is that parallel computing can accelerate the speed of the image generation in the generator block, as well as the real and artificial images generation in the discriminator block. However, the speed of other procedures in training such as propagation and weight updating in the backward computations does not rely on the parallelism improvement. Hence, the speed increase rate becomes extremely slow in large parallelism designs while the area cost increases fast. Based on the results, the computing parallelism size is set to be 32 in the developed memristor-based GAN accelerator.

5.3 Memristor-based GAN Computing Speed

In this section, the speed of the memristor-based GAN accelerator is evaluated and compared with conventional platforms such as GPU and FPGA in previous works. Two big dataset is chosen—ImageNet and Lsun/bedroom is chosen for better demonstration and the dataset detail is listed in Table 3. As is discussed above, the data bitwidth in the proposed memristor-based design is 8 and the parallelism size is 32.

The batch size is set to be 64 in these three scenarios. The speed of GAN training process is evaluated based on the fact that the testing process also works as an inter-step in GAN training.

The experimental results are listed in Table 4. It is observed that our proposed design can achieve $2.8\times$ ($2.7\times$) compared with GPU and $5.5\times$ ($4.8\times$) speedup compared with the FPGA-based accelerator on Lsun (ImageNet). In addition, our design has higher speedup for larger dataset because the parallel pipeline and the memory-free structure can largely decrease the time cost as is discussed in Section 4.

We also analyze the time cost of each computing procedure

Table 3: Dataset

	Dateset Size	Image Size
ImageNet	456567	400×300
Lsun/bedroom	3033042	256×256

Table 4: Speed of the memristor-based, GPU, FPGA-based accelerator

	ImageNet		LSUN/bedroom	
	time(h)	speedup	time(h)	speedup
This Work	6.3	-	47.2	-
GPU	17	$2.7\times$	130	$2.8\times$
FPGA	30	$4.8\times$	255	$5.5\times$

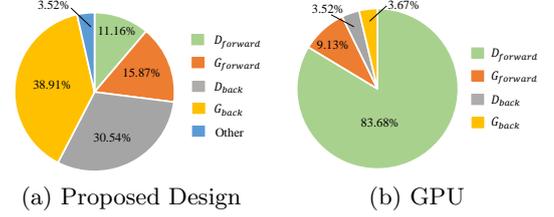


Figure 12: Time cost for each procedure of training GAN

of GAN training that described in Section 2.1, and the results are indicated in Figure 12. As is shown in Figure 12 (b), major time cost in the GPU based GAN computations occurs in the $D_{forward}$ procedure. However, our design reduces such a cost efficiently from 83.68% to 11.16%, as is demonstrated in Figure 12 (a). Our proposed design performs highly resource usage compared with the GPU own to the developed cross-parallel pipeline, hence the GAN computing speed is improved efficiently.

5.4 Energy and Area Cost Analysis

In this section, we evaluate and compare the energy and area cost of the memristor-based GAN accelerator. The energy cost of each procedure in GAN training is also analyzed in detail.

Table 5 shows the energy cost comparison of the memristor-, GPU-, and FPGA-based accelerator. Our proposed design achieves $6.1\times$ and $1.4\times$ energy saving compared with the GPU and FPGA-based GAN computing respectively. The energy cost of each training procedure is analyzed as Figure 12. It is observed that the energy cost of the $D_{forward}$ in the proposed design has a lower energy cost percentage in the whole training process compared with the computations on GPU. Such a low energy cost owns to our proposed memory free data flow in the memristor-based accelerator, in which the energy cost of data communication between memory and CNN (or DeNN) is saved.

The area of the proposed design is $1644mm^2$ when the parallelism size is designed to be 32. The parallel forward flows in the discriminator-block and generator-block accounts for 44.8% and 49.7% respectively. The area of the parallel for-

Table 5: Energy Cost of the ReRAM/GPU/FPGA-based accelerator

	ImageNet		LSUN/bedroom	
	energy (KW/h)	saving	energy (KW/h)	saving
This Work	0.51	-	3.8	-
GPU	3.1	$6.1\times$	23.4	$6.1\times$
FPGA	0.79	$1.5\times$	5.5	$1.4\times$

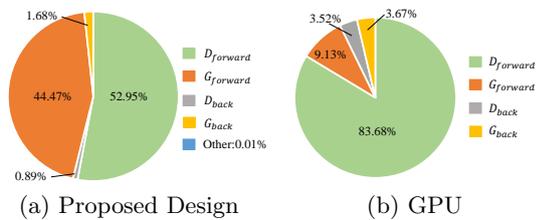


Figure 13: Energy cost for each procedure of training GAN

ward flow is positively related to the parallelism size. When the parallelism size is equal to 1, the area of parallel flows in discriminator and generator-block are equal to $23.0mm^2$ and $25.5mm^2$ respectively and the total area is $139mm^2$.

6. CONCLUSION

Generative adversarial network (GAN) is an effective unsupervised model that is extremely computationally expensive. To address this issue, we proposed a memristor-based accelerator. The proposed design has two major aspects including a cross-parallel pipeline and the memory-free flow. The proposed accelerator was tested on large dataset: ImageNet and Lsun. With area equal to $1644mm^2$, the proposed accelerator can achieve $2.8\times$ ($2.7\times$) and $5.5\times$ ($4.8\times$) speedup, as well as $6.1\times$ ($6.1\times$) and $1.4\times$ ($1.5\times$) energy-saving compared with GPU-based and FPGA-based GAN accelerators respectively on Lsun (ImageNet) dataset.

7. REFERENCES

- [1] R. Girshick, "Fast r-cnn," in *IEEE International Conference on Computer Vision (ICCV)*, pp. 1440–1448, 2015.
- [2] K. He *et al.*, "Deep residual learning for image recognition," in *Advances in Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [3] L. Wang *et al.*, "Visual tracking with fully convolutional networks," in *IEEE International Conference on Computer Vision (ICCV)*, pp. 3119–3127, 2015.
- [4] R. Socher *et al.*, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, pp. 1642–1649, 2013.
- [5] Y. LeCun *et al.*, "Gradient-based learning applied to document recognition," in *Proceedings of the IEEE*, pp. 2278–2324, 2014.
- [6] A. Krizhevsky *et al.*, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the International Conference on Neural Information Processing Systems (NIPS)*, pp. 1097–1105, 2012.
- [7] M. Khanum *et al.*, "A survey on unsupervised machine learning algorithms for automation, classification and maintenance," in *International Journal of Computer Applications*, pp. 0975–8887, 2015.
- [8] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proceedings of the International Conference on Neural Information Processing Systems (NIPS)*, pp. 2672–2680, 2014.
- [9] A. Radford *et al.*, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015.
- [10] W. Zhao *et al.*, "F-cnn: An fpga-based framework for training convolutional neural networks," in *IEEE International Conference on Application-specific Systems, Architectures and Processors (ASAP)*, pp. 107–114, 2016.
- [11] P. Chi *et al.*, "Prime: A novel processing-in-memory architecture for neural network computation in rram-based main memory," in *ACM/IEEE Annual International Symposium on Computer Architecture (ISCA)*, pp. 27–39, 2016.
- [12] L. Song *et al.*, "Pipelayer: A pipelined rram-based accelerator for deep learning," in *IEEE International Symposium on High Performance Computer Architecture (HPCA)*, 2017.
- [13] M. Song *et al.*, "Towards efficient microarchitectural design for accelerating unsupervised gan-based deep learning," in *IEEE International Symposium on High Performance Computer Architecture (HPCA)*, pp. 66–77, 2018.
- [14] J. Ahn *et al.*, "A scalable processing-in-memory accelerator for parallel graph processing," in *ACM/IEEE Annual International Symposium on Computer Architecture (ISCA)*, pp. 105–117, 2015.
- [15] S. Esser *et al.*, "Convolutional networks for fast, energy-efficient neuromorphic computing," in *Proceedings of the National Academy of Sciences of the United States of America*, 2016.
- [16] C. Liu *et al.*, "A spiking neuromorphic design with resistive crossbar," in *ACM/EDAC/IEEE Design Automation Conference (DAC)*, 2015.
- [17] C. Liu *et al.*, "Rescuing memristor-based neuromorphic design with high defects," in *Proceedings of the 54th Annual Design Automation Conference (DAC)*, 2017.
- [18] A. K. Maan *et al.*, "A survey of memristive threshold logic circuits," pp. 1734–1746, 2017.
- [19] E. Lehtonen *et al.*, "Cnn using memristors for neighborhood connections," in *International Workshop on Cellular Nanoscale Networks and their Applications (CNNA 2010)*, pp. 1–4, 2010.
- [20] X. Dong *et al.*, "Nvsim: A circuit-level performance, energy, and area model for emerging nonvolatile memory," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, pp. 994–1007, 2012.
- [21] A. Krizhevsky, "Learning multiple layers of features from tiny images," in *University of Toronto*, pp. 1–60, 05 2012.
- [22] H. A. F. Almurib *et al.*, "A memristor-based lut for fpgas," in *The 9th IEEE International Conference on Nano/Micro Engineered and Molecular Systems (NEMS)*, pp. 448–453, 2014.
- [23] D. Niu *et al.*, "Design trade-offs for high density cross-point resistive memory," in *ACM/IEEE International Symposium on Low Power Electronics and Design (ISLPED)*, 2012.
- [24] D. Querlioz *et al.*, "Simulation of a memristor-based spiking neural network immune to device variations," in *The International Joint Conference on Neural Networks*, pp. 1775–1781, 2011.
- [25] J. Yang *et al.*, "Memristive switching mechanism for metal/oxide/metal nanodevices," pp. 429–433, 2008.
- [26] F. Yu *et al.*, "Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop," *arXiv preprint arXiv:1506.03365*, 2015.