ISSN (Online) 2394-6849

International Journal of Engineering Research in Electronics and Communication Engineering (IJERECE)
Vol 5, Issue 5, May 2018

A Proficient DLAU Design for FPGA Implementation

Vijayashree
SDM College of Engineering & Technolgy Dharwad 580002. Karnataka, India.

Abstract- **These days, the size of systems are increasingly large scale due to the practical applications, which poses significance importance in the field of neural networks. Deep neural networks(DNN)has been employed for image recognition since it can accomplish high exactness by copying conduct of optic nerve in living animal. In order to enhance the execution and additionally to keep up low power cost, in this paper, we design deep learning accelerator unit(DLAU), which is the scalable accelerator for large-scale networks using field-programmable gate array(FPGA) as hardware prototype. In order to improve throughput, it utilizes the tile techniques and employs three pipelined processing units to explore the locality for deep learning applications.**

Keywords: **Deep neural network, tile technique**

I. INTRODUCTION

In recent years, the deep learning has become pervasive in many of the various research fields and commercial applications, and achieved satisfactory products. The demonstration of deep learning accelerator unit (DLAU) on the state-of art Xilinx FPGA board, is used for designing the accelerator unit, which is a scalable accelerator architecture for large-scale neural networks, which improve the performance as well as to maintain the low power cost. The success in the Deep learning is unstoppable, which has speeded up the development of machine learning and artificial intelligence [1].

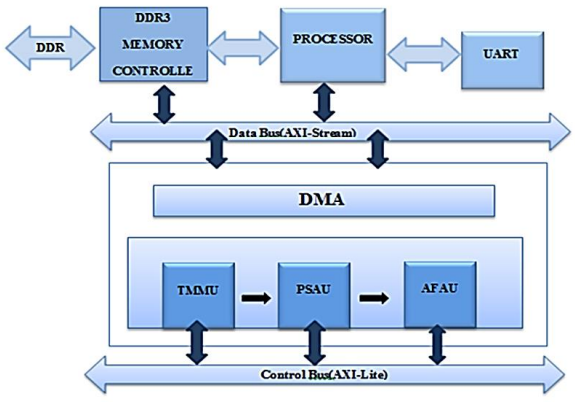

Fig 1: DLAU accelerator architecture

In general, neural networks are multilayered networks, the deep learning uses these multilayered network model to extract high-level abstractions to find the distributed data features in order to solve the complex machine learning problems [2].

II. LITERATURE SURVEY

As designated by chao wang[1],the research field of machine learning, deep learning shows an outstanding capability in resolving complex problems. As the size of the network becomes bulky, it poses extensive challenge to construct a high performance implementation of deep learning neural networks, so as to improve the performance as well as to sustain the low power cost. In this paper it focuses on the implementation of accessible accelerator architecture, DLAU using FPGA as a hardware sample.

D.L.Ly and P.Chew,[2]explain the primary cause for limitation of the neural networks in commercial and industrial applications. The neural networks are typically employed as the software running on general-purpose processors. The algorithms of neural networks that runs in software are typically of O(n2) but the proposed Multi-purpose hardware framework is designed to reduce the O(n2) into an O(n) resources.

As indicated by Chen Zhang,pengLi.jasonCong[3] convolution neural network (CNN) plays a vital role in image recognition, to achieve high accuracy by emulating behavior of optic nerves in living creatures. Recently, rapid growth of modern applications based on deep learning algorithms has further improved research and implementations on FPGA platform because of its advantages of high performance, reconfigurablity, and fast development. But one critical problem is that the computation throughput may not well match the memory bandwidth provided an FPGA platform.

In[4] author Q.YU,C.Wang,X.ma,X,Li and X.Zhou, intimates the emerging field of machine learning, deep

ISSN (Online) 2394-6849

International Journal of Engineering Research in Electronics and Communication Engineering (IJERECE)
Vol 5, Issue 5, May 2018

learning is widely used in applications and cloud services that has the ability in solving complex learning problems. In this paper the design of deep learning accelerator is proposed using FPGA in which accelerator focuses on the implementation of the prediction process, data access optimization and pipeline structure.

After surveying the above mentioned papers one thing is observed that for large-scale network where direct mapping is not possible, the implementation becomes a problem in terms of performances and hardware resources, to tackle this problem, time sharing and reused technology, has been employed that decomposes the input data into data fragments. At every iterations reuse the arithmetic computations to process a data fragments. The proposed work focuses on data access optimization and high pipeline structure.

II. DLAU ARCHITECTURE AND EXCEUTION MODEL

The Fig.1 describes the DLAU accelerator architecture which mainly consists of three pipelined processing units i. TMMU ii. PSAU and iii. AFAU. Each unit has its own functionality. The programming interface and communication between the accelerator and user is provided by means of JTAG-UART. The input data and weight matrix is transferred to the BRAM blocks, from the input buffer. Then the accelerator get activated, once it completes the execution, it returns the results to the user. In order to improve the throughput, it utilizes the tile technique and reads the tiled data by DMA from the memory and after computing with all the three processing units, it writes the result back to the memory.

Some of the key features of DLAU accelerator are listed below;

FIFO Buffer: Buffers are hired to avoid the data loss caused by varying throughput between each processing unit.

Tiled techniques: In order to breakdown the huge volume of data into small slices that can be cached on chip, and to adopt the accelerator to different neural network size tile technique has been employed.

Pipeline accelerator: The computation and data transfer between the neighboring processing units, can be achieved in stream-like manner using pipeline accelerator.

3.1 TMMU Architecture

TMMU is a primary computational unit, whose main functions are multiplication and accumulation. This unit consist of two buffers, an input buffer and output buffer, to receive the input data and to send the computed part sum. It has two registers Reg_a and Reg_b which are used alternately; that reads all the tiled values (set size=32 as an example) .

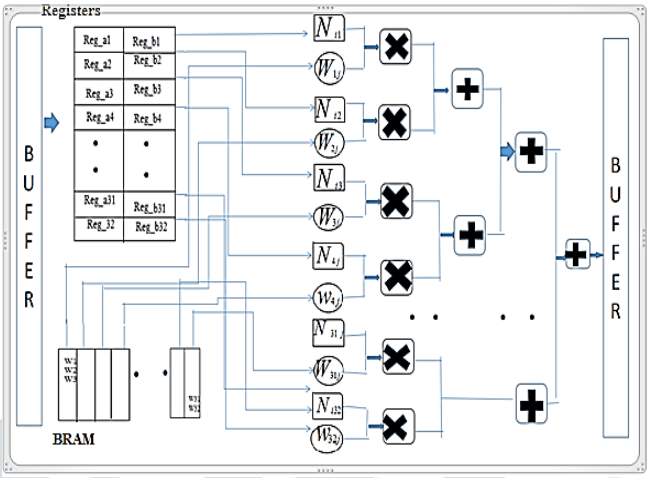

Fig 2.TMMU Schematic

In Fig.2, the BRAM and Registers stores the weight data and node data respectively. A binary tree adder tree like structure is used to optimize the performance.

3.2 PSAU Architecture

Similar to the TMMU, PSAU also writes the result directly into the output buffer, if the part sum produced by the previous unit is the final result. It's main function is to carryout accumulation operation, it can accumulate a single part sum at every cycle. So that it matches the throughput of part sums generated in TMMU.

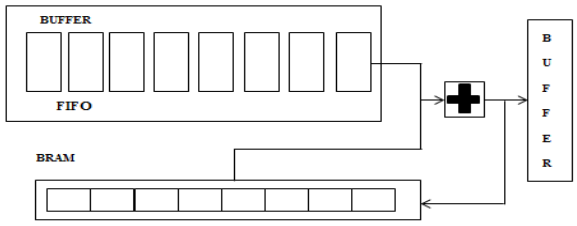

Fig 3: PSAU Schematic

The design of PSAU using PASTA(parallel self-timed adder) shows its superiority in terms of performance and resource utilization over CSA(Carry save Adder). The area can be reduced using PASTA compared to CSA, on the other hand it consumes less hardware resources than CSA.

ISSN (Online) 2394-6849

International Journal of Engineering Research in Electronics and Communication Engineering (IJERECE)
Vol 5, Issue 5, May 2018

3.3 AFAU Architecture

AFAU is also constructed in the similar manner as PSAU in which two separate Block RAMs are employed to store Reg_a and Reg_b values.It is mainly responsible for activation function. It ensure the peak throughput of the accelerator by utilizing fully pipelined architecture.

IV. RESULT ANALYSIS

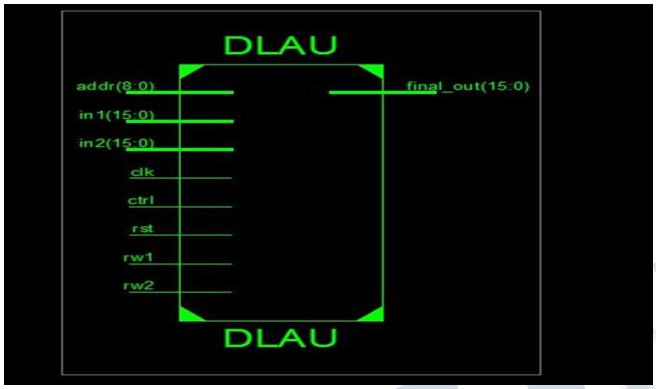

Fig 4 : Block Module of DLAU Accelerator

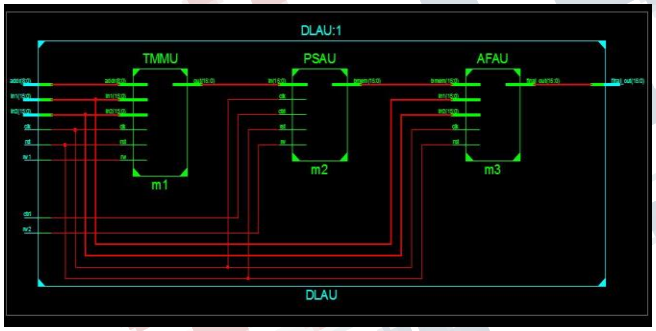

Fig 5: RTL schematic of DLAU Accelerator

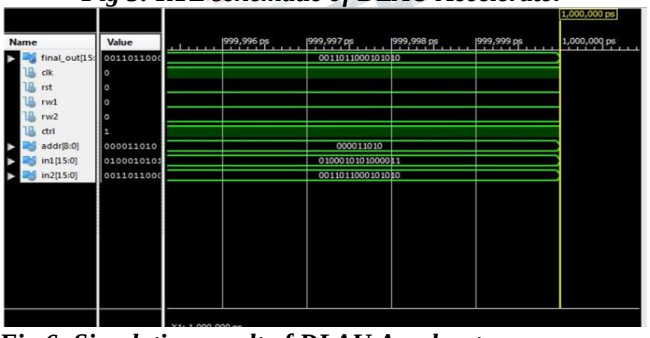

Fig 6: Simulation result of DLAU Accelerator

4.1 Device Utilization Table

Table 1: Resource utilization table of existing DLAU

Name of the component, SPARTAN3E, XC3S250E	SLICES	LUTs	Flip-Flops	IOBs	Power (W)
DLAU (with CSA)	99	84	79	52	0.082

Table 2: Resource utilization table of proposed DLAU

Name of the component, SPARTAN3E, XC3S250E	SLICES	LUTs	Flip-Flops	IOBs	Power (W)
DLAU (with PASTA)	84	57	64	52	0.081

Table 3: Resource utilization of Proposed DLAU

Name of the component, SPARTAN3E,XC3S250E	Logic Utilization	TMMU	PSAU	AFAU
AREA	Number of Slices	16 out of 4656	29 out of 4656	8 out of 4656
	Number of LUTs	-	41 out of 9312	16 out of 9312
	Number of Flip-Flops	32 out of 9312	16 out of 9312	16 out of 9312
	Number of IOBs	50 out of 232	36 out of 232	66 out of 232

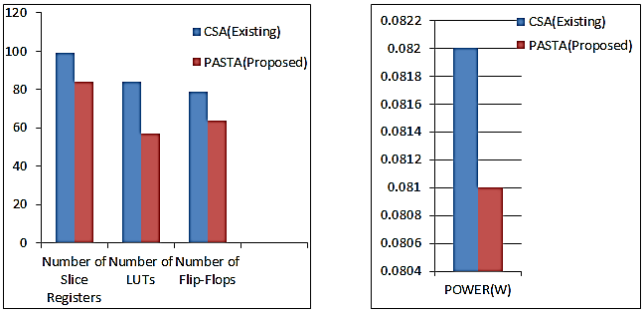

Fig.7: Area and power comparison between Existing and Proposed DLAU

V. CONCLUSION

The proof above demonstrates the significance of accelerator and the idea of deep learning that has

represented ultimately. This innovation works effectively by making utilization of artificial neurons, by extricating the data from the masked layers to take care of the complicated machine learning issues. The adopted architecture is efficient and flexible for large amount of data, implementation of PASTA in place of CSA increases the effiency and reduces the area and power consumption of the accelerator. The experimental results shows that the device utilization and summary of DLAU accelerator that requires a small amount LUT slices, lesser number of bounded IOBs and a smaller amount of IOs evaluate to conventional method. The outline of the accelerator is efficient as far as rate, cost, area and execution, along these lines the throughput is enhanced by diminishing area and power.

VI. ACKNOWLEDGEMENT

I take this opportunity to express my deep sense of greatness to my guide Mr. M. Vijay Kumar, Asst. Professor, Dept. of E&C Engg., for his valuable advice, expert guidance and support at all stages during the course of project. I acknowledge my sincere thanks to the principal, SDMCET, Dharwad. I would also like to thank to Dr.G.A Bidkar,HOD.,Dept. of E&C Engg., and our P.G. Cordinator,M.Vijaya.C for their support and I would like to extend my thanks to all the faculty members and well wishers for their timely help to carry out this project.

REFERENCES

1. Chao Wang,Li Gong,Qi yu,Xi Li, "A Scalable Deep Learning Accelerator Unit on FPGA", vol.36, No.3,2017.

2. D.L.Ly and P.Chow,"A high-performance FPGA architecture for restricted Boltzmann machines", in proc. FPGA, Montery, CA, USA, 2009, pp,73-82.

3. C.Zhang et al.,"optimizing FPGA-based accelerator design for deep convolution neural networks", in proc. FPGA, Monterey, CA, USA, 2015

4. Q.Yu, C.Wang,X.Ma, X.Li, and X. Zhou, "A Deep Learning Prediction process accelerator based FPGA", in proc. CCGRID, Shenzhen, china, 2015, pp.1159-1162.

5. T. chen et al., "DianNao: A small-footprint high-throughput accelerator for obiquitous machine –learning," in proc.ASPLOS,salt Lake city, UT,USA,2014,pp.269-284

6. S.K.Kim,L.C. McAee, P.L. McMahon, and K. Olukotun, "A highly scalable restricted Boltzmann Machine FPGA implementation", in proc. FPL,prague,Czech Republic,2009,pp.367-372

7. J.Qu et al., "Going deeper with embedded FPGA platform for convolution neural network", in proc. FPGS, Montery, CA,USA, 2016, pp.26-35.

8. P. Thinbodeau. Data centers are the New polluters. Accessed on Apr. 4, 2016.[online]. Available: http://www.computerworld.com/ article/2598562/data-center/data-centers-the-new-polluters.html

9. J. Hauswald et al., "DjiNN and Tonic: DNN as a service and its implictions for future warehouse scale computers." In proc. ISCA, Portland. OR. USA, 2015, pp,27-40

10. P. Ferreira, P. Ribera, A. Attunes, and F. M. Dias, "A high bit resolution FPGA implementation of FNN with a new algorithm for activation of FNN with a new algorithm for the activation function", Neurocomputing, vol.71,pp.71-77,2007.

11. K. Simonyan and A. Zisserman, "very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1408.1556,2014.

12.D. Liu, T. Chen, S. Liu, J. Zhou, S. Zhou, O. Teman, X. Feng, X. Zhou, and Y. Chen, "pudiannao: A polyvalent machine learning accelerator," in ASPLOS ACM, 2015, pp 369- 381.

13.M. A. Erdogdu, "Newton-stein method: A second order method for g1ms via steins lemma", in advances in neural Information processing systems 28, 2015, pp. 1216-1224..