A Power Effective DLA for PBs in Opto-Electrical Neural Network Architecture

Ralph Gerard B. Sangalang[†], Shih-Heng Luo[†], Hsin-Che Wu[†], Bao-Qi He[‡], Shen-Fu Hsiao[‡],
Chua-Chin Wang[†]*, Chewnpu Jou[§], Harry Hsia[§], and Douglas C.-H. Yu[§]
[†]Dept. of Electrical Engineering, National Sun Yat-Sen University, Kaohsiung, Taiwan 80424
[‡]Dept. of Computer Science and Engineering, National Sun Yat-Sen University, Kaohsiung, Taiwan 80424
[§]Taiwan Semiconductor Manufacturing Company, Hsinchu, Taiwan 30078

***Abstract*—Deep neural networks (DNN) have been widely used in many real-time artificial intelligent (AI) applications because of effective hardware accelerators. However, most present designs either suffer from high area cost or low hardware usage. This paper presents a design of a digital logic accelerator (DLA) for use in PBs (processing block) of an opto-electrical neural network (OENN). The proposed DLA uses processing elements that detects underflow and overflow. Besides, it also increased the processing time to resolve the timing problems. The details of the design together with post-layout simulations are presented in this paper. The DLA is implemented using a typical 40-nm CMOS process. It showed a performance result of 51.2 GOPS and the power consumption is 91.3 mW at 125 MHz.**

***Index Terms*—deep neural networks (DNN), hardware accelerators, deep learning, energy-efficient accelerators, opto-electrical integration.**

I. Introduction

Deep learning has prospered in recent years owing to its capacity to discover patterns within data and thereby pave the way for intelligent decision making, which is superior in certain situations to human capabilities. At the moment, neural networks utilizing electronic systems had applications that greatly benefited the field of sound processing, video processing, communication systems, pattern analysis, etc. In the center of these neural network applications are convolutional neural networks (CNN) and deep neural networks (DNN) inspired by extracting features in a small region of a specific visual specimen to attain patterns in other regions of the specimen. By using optimized algorithms, detection and prediction are done with high accuracy and converges extremely fast. Neural networks using electronic systems are using power hungry hardware such as CPU, CPU, and FPGA.

Electronic neural networks (ENN) have been proven to have many useful applications but still is limited by the drawbacks of transferring electrons for signal processing and transmission. ENNs suffered from a high time delay due to the fact that massive weight coefficients needed to be transferred from memory modules to processing units, and then transmits the results back again to the memory unit. This issue results in poor power efficiency. Moreover, as the neural network becomes larger, this data transfer needs to be faster which makes the devices harder to catch up and manufacture.

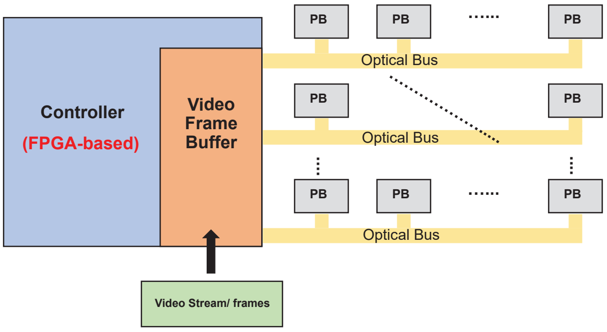


Fig. 1. Conceptual Opto-Electrical Neural Network (OENN) architecture.

To improve upon the transfer rate problem, photons have been used as the carrier of the information signals through optical neural networks (ONN). Reck *et al.* presented the most important design for chip-integrated optical neural networks in 1994 [1]. ONNs has been investigated to overcome such problems in time delays compred to ENN. Data transfers between layers and modules can be done at the speed of light using high speed fiber optic cables. Hence, combining the capabilities of an electronic and optical neural networks was considered as a feasible solution [2]. Fig. 1 is a conceptual OENN architecture, where a distributive array of PBs (processing blocks) are connected to an optical bus. The optical bus was governed by a controller with a data buffer. When the video streams are read, every frame will be transferred over the optical bus to a designated PB to be processed.

This paper presents a design of a digital logic accelerator (DLA) to be used in speeding up the processing speed of an Opto-Electrical Neural Network (OENN). The design presents a new processing element that operates faster than prior designs. The design is implemented using a typical 40-nm CMOS technology. Post-simulations are done to show the performance of the DLA.

II. Design of the DLA

A. Hardware Architecture

Referring to Fig. 2, the proposed DLA composed of a DNN Accelerator, an Inter-controller, and a AXI Wrapper Direct

*Dr. Chua-Chin Wang is the correspondence. Dr. Wang is also with Inst. of Undersea Technology, National Sun Yat-Sen University, Taiwan. (Email: ccwang@ee.nsysu.edu.tw) This investigation was supported by Taiwan Semiconductor Manufacturing Corporation Limited, Taiwan under TSMC Contract no. 202201-100011.

978-1-6654-5071-3/22/$31.00 ©2022 IEEE

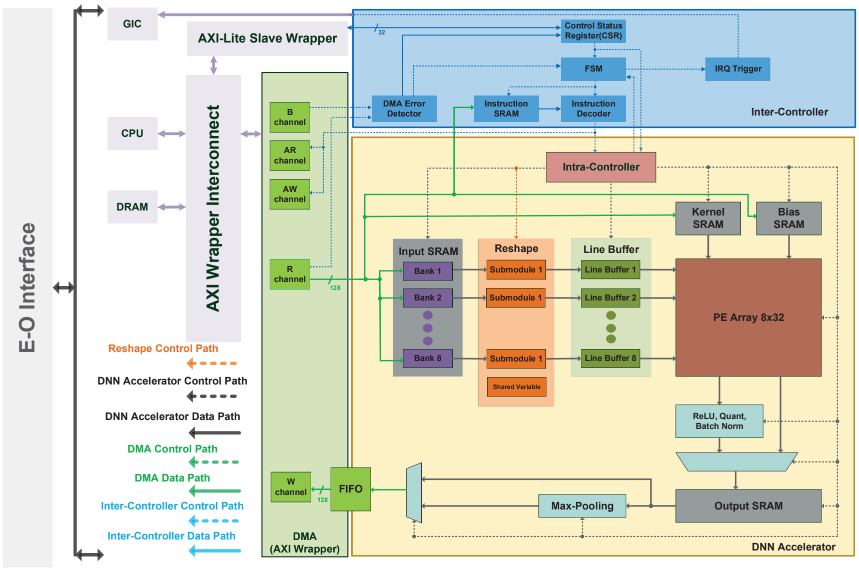


Fig. 2. Proposed DLA in a PB.

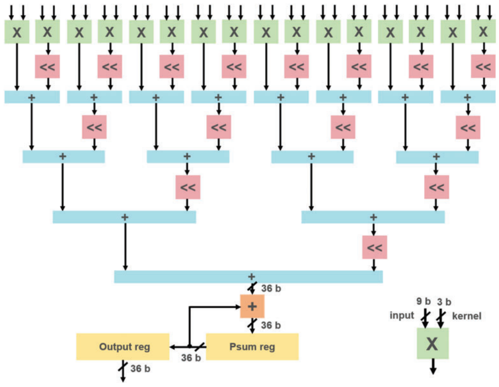


Fig. 3. Prior processing element (PE) design [3].

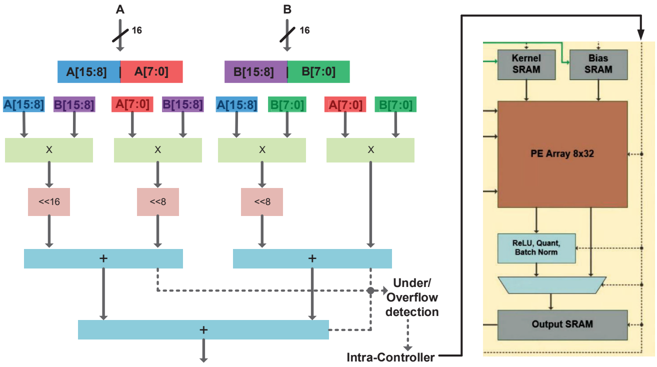


Fig. 4. Proposed processing element (PE) with overflow and underflow detector.

Memory Access (DMA) is shown. The DNN Accelerator consists of an 8×32 PE array, an intra-controller, SRAMs, reshape modules, and line buffers.

The convolution operation can be realized by three parallel computing methods, namely 1.) Input Channel Parallel (ICP), 2.) Output Channel Parallel (OCP), and 3.) Window Parallel (WP). These methods provides different memory allocation structure that aids in the power usage and performance enhancement.

B. *Proposed Processing Element (PE)*

The processing element reported in [3] is a multi-resolution architecture as shown in Fig. 3. It can handle different weight resolution for the operations. Although it offers an advantage of a multi-resolution ability, the trade-off is a longer delay because of the larger number of stages. Another penalty of this design is when overflow or underflow happens, it just forwards the result to the next stage hence causing errors in calculations.

To resolve the delay and inevitable overflow & underflow problems, a new processing element was presented as shown in Fig. 4. Instead of a multi-resolution architecture, the proposed PE element will only use 16-bit resolution without loss of robustness. It is composed of four 8×8 multipliers (as shown in Fig. 5), one 16-bit logical shifter, two 8-bit logical shifters, three 16-bit adders, and a underflow/overflow detector. By this, the number of shifting steps are drastically reduced because the element no longer need to create trailing bits for the inputs of the PE. It also has less number of stages compared to the prior designs, thus offering better speed performance. In addition, an underflow and an overflow detector is introduced to the new PE to co-work with the intra-controller and send signals to the quantization module of the DLA. Overflow is detected when the result is over 16’h7FFF and underflow when result is 16’h8000.

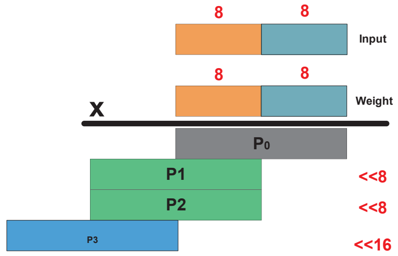


Fig. 5. 8×8 multiplier calculation in the PE.

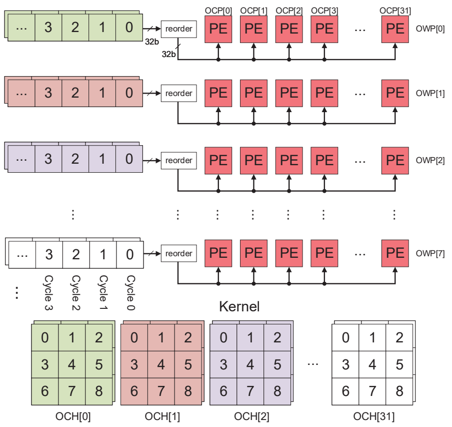


Fig. 6. PE array architecture.

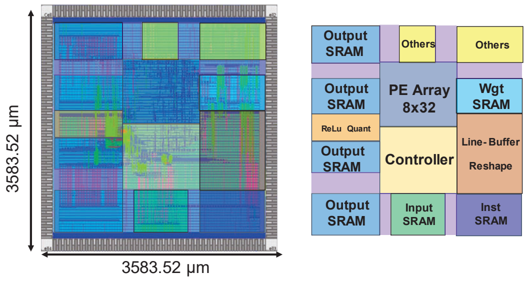


Fig. 7. Layout of the hardware accelerator.

C. PE Array

Fig. 6 shows the PE array architecture composed of 8 rows of output window parallel (OWP) and 32 columns of output channel parallel (OCP). The outputs are all passed to the left and a set of kernel values are shared by the same column. To reduce the area and power consumption of the array, the excitation functions, quantization, and batch normalization of the pixels are performed outside the array prior to the output SRAM.

D. Input, Kernel, and Output SRAM

The input and kernel SRAM of the proposed design uses 8 banks 2-port SRAMs (1R1W)s of 8 bits width. The complete bank is made of double buffers to shorten the access time. Another advantage of the double buffer design is that the resource can be allocated to other operation when not in need. That is, when only half of the bank is used, the other half can be allocated to load the required data. The output SRAM of the proposed design uses 4 banks of 128 bits wide dual-port SRAM (2R or 2W or 1R1W). The line buffer in the proposed design uses 16-bit registers with 128-bit SRAM (1R1W) and pushes 8 groups of 16-bit data during the convolution operation.

E. Reshape and Line Buffer Modules

A reshape module is used for the design to support tile-based calculations. The reshape module reorganizes the tile based data to support burst transmission to make it possible to transmit multi-tile data from one module to another. It uses a padding methods that reorganizes the data into 1D or 2D representation prior to transmission.

Line buffers are also used to hold the weights and data values that will be repeated during convolution operations. It uses a 16-bit register matching the width of the input SRAM.

F. Controller

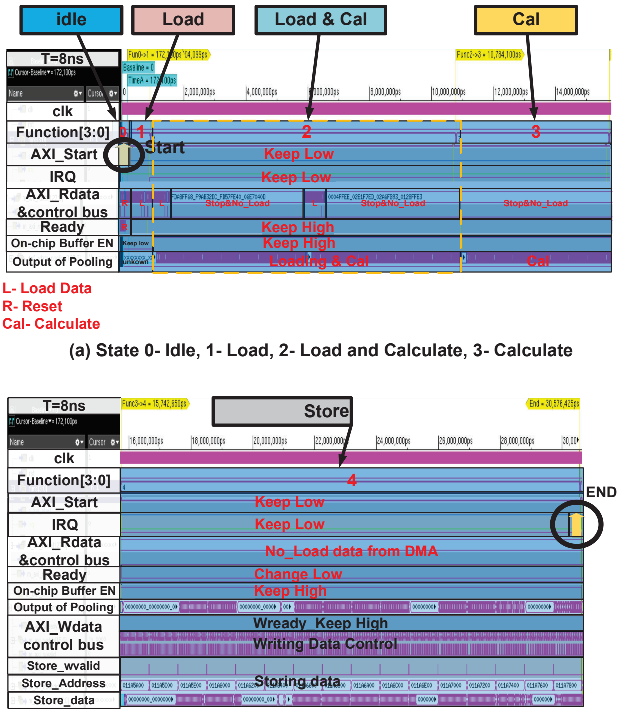
The controller uses the Inter-Controller to control the DMA and all the other hardware. The control signals determine the operation of the connected modules, including the finite-sate machine (FSM) state transfer, circuit switching, and power consumption. The Inter-Controller has five states: 1. Idle (S0) state, 2. Load (S1) state, 3. Calculate & Load (S2) state, 4. Calculate (S3) state, and 5. Store (S4) state.

III. Implementation and Verification

The proposed digital logic accelerator is implemented using TSMC 40-nm CMOS technology. Fig. 7 shows the layout of the chip. The DLA has a size of 3583×3583 μm^2 including pads. Fig. 8 shows post-layout simulations in the worst case, namely slow-slow (SS) corner, at a frequency of 125 MHz. It shows results consistent with the pre-layout simulations.

To verify the DLA functionality, two sets of results were prepared: 1) from CPU-based software simulations and 2) the DLA outcome. An algorithm based on YoloV3-tiny was implemented with both solutions and then are compared with each other to estimate the absolute error. The absolute error is found within 1.4% as shown in Fig. 9. Fig. 10 shows sample results of the DLA solution with CPU solution applied to object recognition of underwater objects using a YoloV3-tiny-based algorithm. We also included two scan chains and BIST (using March algorithm) to enhance the DLA's reliability. The test coverage was up to 98.5%.

Table I shows the comparison with many recent NN accelerator works. Notably the supply voltage of our DLA is 0.9 V operating at 125 MHz frequency. The simulation results show a performance 51.2 GOPS at a power consumption of 91.3 mW. The proposed design shows an FOM value of 63.15 which is the best among all works. In other words, TOPS/W = 0.561, and GOPS/mm^2 = 3.99, both are the best by far if normalized with CMOS technology nodes. It also shows the



(a) State 0- Idle, 1- Load, 2- Load and Calculate, 3- Calculate

(b) State 4- Store

Fig. 8. Worst-case post-layout simulation.

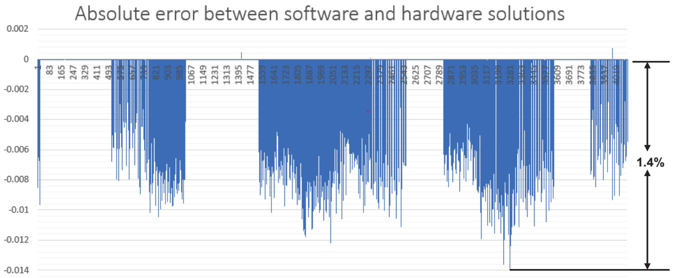


Fig. 9. Absolute error between hardware and software solutions (max. error = 1.4%).

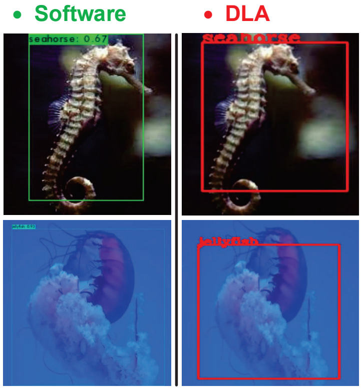


Fig. 10. Comparison of CPU-based and DLA-based simulations for detecting underwater objects.

TABLE I
COMPARISON TABLE

	[4]	[5]	[3]	Ours
Year	2018	2020	2022	2022
Process (nm)	65	40	180	40
Verification	Simu.	Simu.	Meas.	Simu.
Supply (V)	1.2	0.9	1.8	0.9
Area (mm^2)	10.6	200	53.63	12.84
Max. Freq. (MHz)	200	200	100	125
On-chip buffer (kb)	139.6	118	150	150
No. of MACs	64	128	256	256
Activation bit-width	16	16	16	16
Kernel bit-width	16	16	16	16
Performance (GOPS)	23.4	51.2	40.96	**51.2**
Power (mW)	93.4	153.94	196.8	**91.3**
Area eff. (GOPS/mm^2)	2.21	0.26	0.76	**3.99**
Power eff. (TOPS/W)	0.253	0.3326	0.208	**0.561**
^{1}CO$_2$ equivalent (kg.)	0.35	0.58	0.75	**0.34**
^{2}FOM	60.1	59.86	46.8	**63.15**

^{1}Based on U.S. EPA greenhouse gas equivalency [6]. Computed based on continuous operation for 1 year.

^{2}FOM $= \frac{\text{Frequency(MHz)}\times\text{GOPS}}{\text{Normalized Power(mW)}}$

lowest carbon dioxide (CO$_2$) equivalent enegy emission when used continuously for an entire year.

IV. Conclusion

A low-power and high performance digital logic accelerator using 40-nm CMOS process is presented in this investigation. A new processing element with underflow and overflow detection is proposed to increase the processing speed and reduce computational errors. The simulated performance was found to be 51.2 GOPS at 91.3 mW power consumption with a clock rate of 125 MHz. The carbon dioxide equivalent shows that our design is the most environmentally friendly in terms of energy consumption. The FOM shows that our design is the best so far.

Acknowledgment

The authors would like to thank the Taiwan Semiconductor Manufacturing Corporation for supporting this project.

References

[1] M. Reck, A. Zeilinger, H. J. Bernstein, and P. Bertani, "Experimental realization of any discrete unitary operator," *Physical Rev. Lett.*, vol. 73, no. 1, pp. 58–61, Jul. 1994.

[2] J. Chang, V. Sitzmann, X. Dun, W. Heidrich, and G. Wetzstein, "Hybrid optical-electronic convolutional neural networks with optimized diffractive optics for image classification," *Scientific Rep.*, vol. 8, no. 1, pp. 1–10, Aug. 2018.

[3] C.-C. Wang, R. G. B. Sangalang, C.-P. Kuo, H.-C. Wu, Y. Hsu, S.-F. Hsiao, and C.-H. Yeh, "A 40.96-GOPS 196.8-mW digital logic accelerator used in DNN for underwater object recognition," *IEEE Trans. Circuits Syst. I-Regul. Pap.*, 2022.

[4] J. Jo, S. Kim, and I.-C. Park, "Energy-efficient convolution architecture based on rescheduled dataflow," *IEEE Trans. Circuits Syst. I-Regul. Pap.*, vol. 65, no. 12, pp. 4196–4207, Dec. 2018.

[5] S.-F. Hsiao, K.-C. Chen, C.-C. Lin, H.-J. Chang, and B.-C. Tsai, "Design of a sparsity-aware reconfigurable deep learning accelerator supporting various types of operations," *IEEE J. Emer. Select. Top. Circu. Syst.*, vol. 10, no. 3, pp. 376–387, Sep. 2020.

[6] United States Environment Protection Agency, "Greenhouse gas equivalencies calculator," [Online], Mar. 2022. [Online]. Available: https://www.epa.gov/energy/greenhouse-gas-equivalencies-calculator