2024 IEEE 6th International Conference on AI Circuits and Systems (AICAS) | 979-8-3503-8363-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/AICAS59952.2024.10595977

A 54.61-GOPS 96.35-mW Digital Logic Accelerator For Underwater Object Recognition DNN Using 40-nm CMOS Process

Chua-Chin Wang[†*], Shih-Heng Luo[†], Hsin-Che Wu[†], Ralph Gerard B. Sangalang[†‡], Chewn-Pu Jou[§], Harry Hsia[§], and Lan-Chou Cho[§]
[†]Dept. of Electrical Engineering, National Sun Yat-Sen University, Kaohsiung, Taiwan 80424
[‡]Dept. of Electronics Engineering, Batangas State University, Philippines 4000
[§]Taiwan Semiconductor Manufacturing Company, Hsinchu, Taiwan 30078

***Abstract*—CNN (convolution neural network) and DNN (deep neural network) have been widely used in real-time artificial intelligent (AI) applications, particularly image or video recognitions, because they have been proved physically in many occasions. However, most prior AI hardware works are either suffered from high on-silicon area cost or low usage thereof. This paper presents a power efficient and high performance implementation of a digital logic accelerator (DLA) for the usage of real-time underwater object recognition. The proposed DLA is also featured with 2-dimentional PE (processing element) array to increase the processing throughput by the enhancement of parallelism. The design was realized and fabricated using TSMC 40-nm CMOS process. Not only the post-layout simulation results are shown, the on-silicon measurement outcome is also demonstrated to prove the function correctness and the performance. The area efficiency (GOPS/mm^2) is 4.562, and the power efficiency (TOPS/W) is 0.5668 on silicon, which both are the best to date.**

***Index Terms*—deep neural networks (DNN), digital logic accelerator (DLA), deep learning, chip measurement, high degree of parallelism**

I. Introduction

Artificial intelligence (AI) has been booming in the past years owing to its capacity to discover unnoticed patterns and links within data sets such that it paves the way for data fusion and decision making, which is considered to be superior in certain occasions to human capabilities.

As pointed out earlier, one of the biggest challenges in AI-related applications is to equip AI hardware into applications where the power/energy source is limited, such as battery-powered autonomous underwater vehicles (AUV). That is, the overall power in these underwater applications is constrained by the battery of the vehicle systems. Besides the development of light-weight training algorithms, another approach is to seek more power-efficient hardware solutions to overcome the mentioned power limitation issues for these battery-driven AI systems.

[*]Prof. Chua-Chin Wang is the correspondence. He 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.

Many researchers still paid their attention to existing CMOS-based electronic solution such that many AI hardware accelerators have been reported in past years. The systolic architecture has reconfigurable designs for different convolution kernels [1] [2]. However, a large area overhead becomes the cost to pay due to the complexity of the control structure is found [1]. The systolic data flow is only allowed to run in the PE rows rather than the entire 2D plane to achieve data reuse of filters along with convolution kernel reuse, which results in throughput degradation [2]. Another hardware design is the spatial architecture reported in [3], which is an improved version of [4]. It mainly took advantage of clustering of sparse CNNs to achieve higher throughput, but it required a larger area. A filter and input reuse in the streaming architecture was reported to be energy efficient, but it has a low hardware utilization [5]. More importantly, none of the above works were focused on power-limited underwater object recognition applications. A novel DLA featured with high degree of hardware parallelism is disclosed in this investigation to resolved the mentioned issues.

II. Low-power and High-throughput DLA Design

Fig. 1 shows the required interface for the proposed DLA to co-work with auxiliary circuits. Besides the AXI wrapper, the control and data paths of accelerator, DMA, and Inter-Controller are also defined to drive the PLA. More specifically, the circuitry in the green dashed area is meant to serve as the wrapper I/Os, where many channels are needed to carry out the AXI protocol. The blue dashed area mainly consisting of controller circuitry in charge of the state transition of a FSM (finite state machine) generating all the commands of data flow control, instruction control, and timing control. Last but not least, the yellow area is the core of DLA where all the convolution operations, MAC computation and management, and realization of parallelism are realized.

A. DLA Hardware Architecture

Referring to Fig. 1 again, the proposed DLA is mainly composed of a DNN Accelerator, an Inter-controller, and a AXI Wrapper Direct Memory Access (DMA). Most importantly,

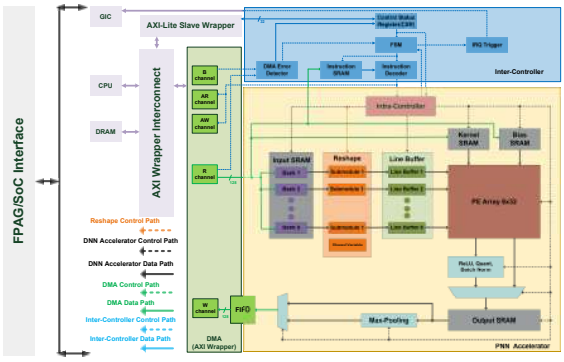

Fig. 1. Interface of the proposed DLA with auxillary units

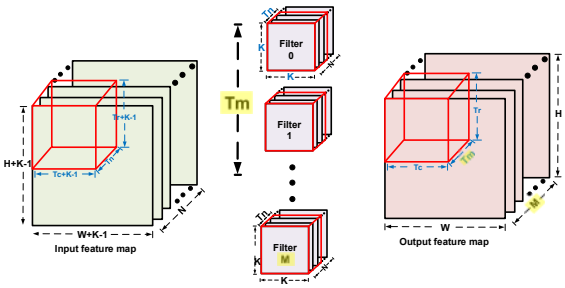

Fig. 2. Core of convolution calculations

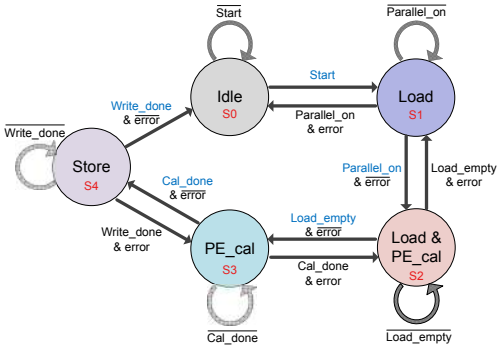

Fig. 3. State transition diagram of the DLA controller.

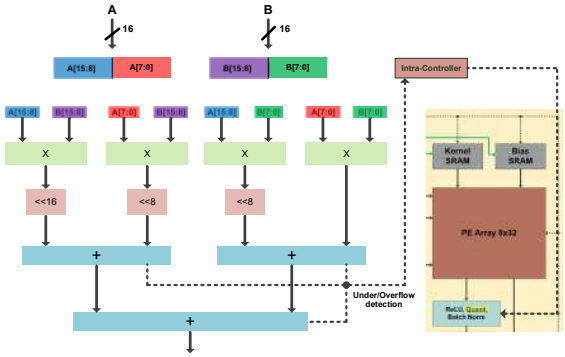

Fig. 4. Proposed processing element (PE) with overflow and underflow detector.

the DNN Accelerator consists of an 8×32 PE array to increase the degree of parallelism, an intra-controller, SRAMs, reshape modules, and line buffers. Apparently, the core of DNN or CNN is the convolution operation, namely MAC operations. To elevate the throughput without serious power overhead, the convolution operation is realized by three parallel computing steps as follows, 1.) Input feature map (IM), 2.) Input Channel Parallel (ICP), and 3.) Output Channels (OCH). The above computing steps are shown in Fig. 2, where H is the frame length, W is the frame width, and K is the size of the Filters.

B. Controller

The DLA controller relies on the Inter-Controller to generate timing signals to drive the DMA and all the other hardware modules. In other words, the controller is the main driver of DLA to generate control signals for the operation of the modules to be activated or not, including the finite-state 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, as shown in Fig. 3. Each state is assigned to carry out different modes by generating corresponding signals.

C. Proposed Processing Element

Most of the prior processing element (PE) was based on a multi-resolution architecture, e.g., [6], which was for different weight resolutions used in MAC operations. Though it attained an advantage of diversity, the trade-off is long delay caused by the cascaded stages. A very fatal issue, on top of the delay, is that the overflow or underflow can not be detected or fixed, if they appear during the MAC operations. Namely, the error will be accumulated and passed downstream if it happens.

To resolve the long delay and inevitable overflow & underflow problems, a new processing element was presented as shown in Fig. 4, where the proposed PE element will only use 16-b resolution integer without loss of robustness for underwater object recognition. It is composed of four 8×8 multipliers, one 16-b logical shifter, two 8-b logical shifters, three 16-b adders, and an underflow/overflow detector. The number of shifting steps are drastically reduced by such an architecture, because the element no longer needs to generate trailing bits for the inputs of the PE. Moreover, less number of stages is needed compared to the prior designs, thus reducing the delay. An underflow and an overflow detector are also introduced to co-work with the intra-controller and send notification signals to the quantization module of the DLA. More specifically, Overflow is detected when the output result is over 16'h7FFF and underflow when it is 16'h8000.

D. PE Array

Fig. 5 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 generated and passed to the left and a set of kernel values are shared by the same column. To reduce the area and power consumption

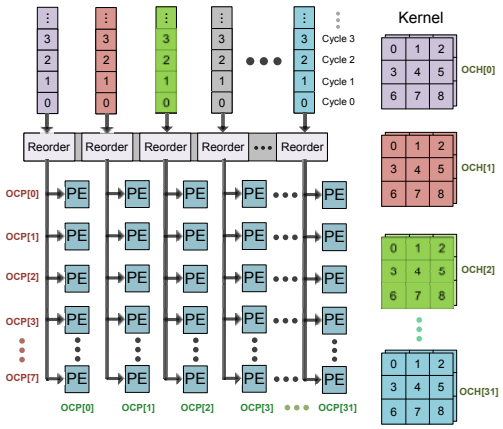

Fig. 5. PE array architecture.

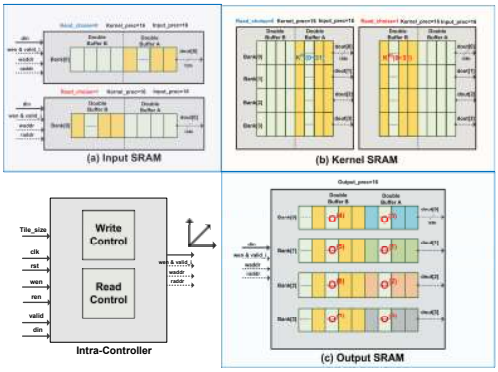

Fig. 6. SRAM structures of (a) Input, (b) Kernel, and (c) Output.

of the array, the excitation functions, quantization, and batch normalization of the pixels are performed outside the array prior to the output SRAM.

E. Memory Arrangement for Input, Kernel, and Output

The input and kernel SRAMs of the proposed DLA was realized with 8 banks 2-port 8-b 1R1W SRAMs, respectively. Double buffers are employed at the outputs thereof to reduce the access time and increase the slew rate. The double buffer design also attains another advantage, which is that the resource can be allocated to other operations when it is not in use. For instance, if only half of the bank is used, the other half is allocated to load the required data. By contrast, the output SRAM of the DLA uses 4 banks of 128-b dual-port SRAMs (2R or 2W or 1R1W). The SRAMs for Input, Kernel, and Output are shown in Fig. 6, respectively.

F. Reshape and Line Buffer Modules

A Reshape module is used for the design to support tile-based calculations. The Reshape module re-organizes the tile-based data to support burst transmission such that the transmission of multi-tile data from one module to another is feasible, as shown in Fig. 7. Notably, a padding step is needed that re-organizes the data into 1D or 2D representation prior to the transmission.

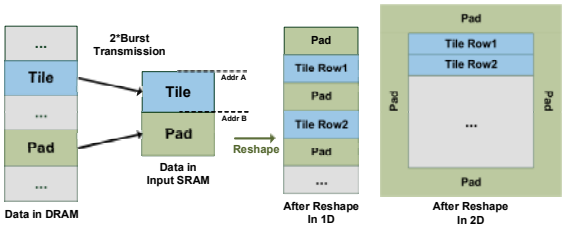

Fig. 7. Opration flow of re-oranization of tile and pad.

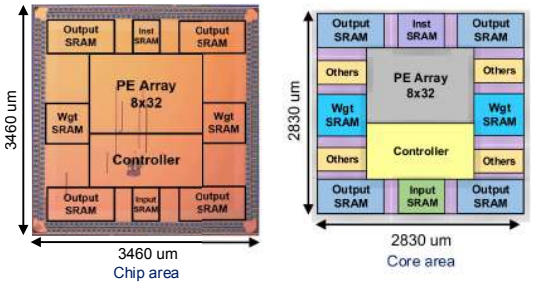

Fig. 8. Diephoto and layout of the DLA.

Lastly, the line buffer is composed of 16-b registers with 128-b SRAM (1R1W), which is used to push 8 groups of 16-b data during the convolution operation. The line buffers are also used to hold the weights and data values, which are repeatedly used during convolution operations.

III. Implementation and Measurement

The proposed digital logic accelerator is implemented using TSMC 40-nm CMOS technology. Fig. 8 shows the diephoto and the layout of the DLA chip, where the on-silicon size is 3460×3460 μm^2 including 320 I/O and power pads. Notably, two scan chains and BIST (using March algorithm) are included to enhance the DLA's reliability and testability, where the test coverage was up to 98.35%.

A. Chip Measurement

The performance measurement of the proposed DLA chip was conducted in SOC Lab. of TSRI in Hsinchu, Taiwan, using ADVANTEST 93000, as shown in Fig. 9. Fig. 10 is the screenshot of the measured waveforms, which match the post-layout simulation result. Lastly, Fig. 11 is the Shmoo plot to show that when VDD = 0.9 V, the operation clock = 133.33 MHz.

B. Function Verification

To verify the DLA functionality, the best way is to make a comparison of two experiments: 1) from CPU-based software simulations (using float32 computation, assumed as the golden model), and 2) the DLA+FPGA hardware testing. An algorithm based on YOLOv3-tiny was implemented using the mentioned two approaches, which are compared to estimate the absolute error caused by our DLA. The comparison experiment was shown in Fig. 12, which shows the identical recognition

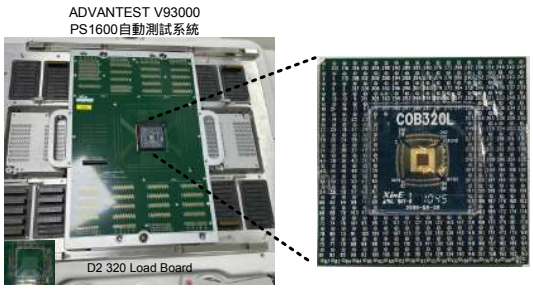

Fig. 9. DLA chip measurement setup.

Fig. 10. Measured waveforms (partial).

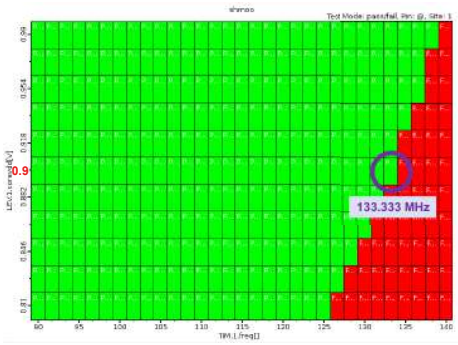

Fig. 11. Shmoo plot of the proposed DLA.

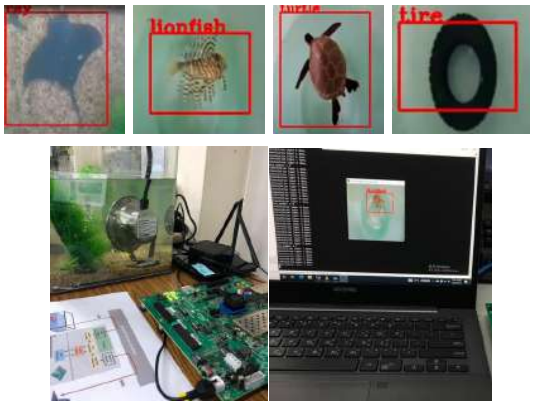

Fig. 12. Comparison of CPU-based and DLA-based experiments for underwater object recognition.

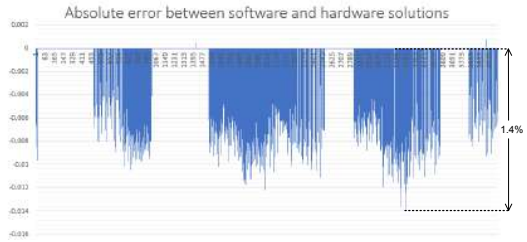

Fig. 13. Absolute error of the DLA approach vs. the golden model ($\leq$ 1.4%).

TABLE I
COMPARISON TABLE

	[4]	[7]	[6]	Ours
Year	2017	2020	2022	2023
Publucation	JSSC	JSSC	TCAS-I	
Process (nm)	65	65	180	40
Verification	Meas.	Meas.	Meas.	Meas.
Supply (V)	1.0	0.6	1.8	0.9
Area (mm^2)	16	2.56	53.63	11.97
Max. Freq. (MHz)	200	0.25	100	133.33
No. of MACs	168	512	256	256
Performance (GOPS)	42	0.471	40.96	**54.61**
Power (mW)	278	0.0106	196.8	**96.35**
Area eff. (GOPS/mm^2)	2.625	0.184	0.7638	**4.562**
Power eff. (TOPS/W)	0.1511	44.434	0.2081	**0.5668**
[1]CO_2 equivalent (kg.)	1.05	0.01	0.75	**0.365**
[2]FOM	30.21	6.40	37.46	**68.01**

[1]Based on U.S. EPA greenhouse gas equivalency [8]. Computed based on continuous operation for 1 year.
[2]$\text{FOM} = \frac{\text{Frequency(MHz)} \times \text{GOPS}}{\text{Normalized Power(mW)}}$

results except the delay and frame rate. The absolute error was found to be less than 1.4% as shown in Fig. 13.

Table I shows the comparison with many recent CNN/DNN hardware accelerator works reported in top journals recently. Notably, the supply voltage of our DLA is 0.9 V operating at 133.33 MHz frequency. The on-silicon measurement results of our DLA show a performance 54.61 GOPS at a power consumption of 96.35 mW. Meanwhile, TOPS/W = 0.5668, and GOPS/mm^2 = 4.562, both are the best by far if normalized with CMOS technology nodes and the operating clock frequency. In short, the proposed design shows an FOM value of 68.01 which is the best among all DLA works. Lastly, it also shows the lowest carbon dioxide (CO_2) equivalent energy emission when used continuously for an entire year.

IV. Conclusion

A low-power and high performance DLA using 40-nm CMOS process is presented in this investigation. A new parallel architecture based on processing element with underflow and overflow detection is proposed to increase processing speed and reduce computational error. Not only the normalized area and power efficiencies of our design are better than prior DLAs, the FOM also shows that our design is the best so far even if the clock frequency is taken into account.

REFERENCES

[1] F. Tu, S. Yin, P. Ouyang, S. Tang, L. Liu, and S. Wei, "Deep convolutional neural network architecture with reconfigurable computation patterns," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.,* vol. 25, no. 8, pp. 2220–2233, Aug. 2017.

[2] Y. Huan, J. Xu, L. Zheng, H. Tenhunen, and Z. Zou, "A 3D tiled low power accelerator for convolutional neural network," in *Proc. IEEE Int.Symp. Circuits Syst. (ISCAS),* May 2018, pp. 1–5.

[3] Y.-H. Chen, T.-J. Yang, J. Emer, and V. Sze, "Eyeriss v2: A flexible accelerator for emerging deep neural networks on mobile devices," *IEEE J. Emerging Sel. Topics Circuits Syst.,* vol. 9, no. 2, pp. 292–308, Jun. 2019.

[4] Y. Chen, T. Krishna, J. Emer, and V. Sze, "Eyeriss: An energy efficient reconfigurable accelerator for deep convolutional neural networks," *IEEE J. Solid-State Circuits,* vol. 52, no. 1, pp. 127–138, Jan. 2017.

[5] L. Du, Y. Du, Y. Li, J. Su, Y.-C. Kuan, C.-C. Liu, and M.-C. F. Chang, "A reconfigurable streaming deep convolutional neural network accelerator for Internet of Things," *IEEE Trans. Circuits Syst. I,Reg. Papers,* vol. 65, no. 1, pp. 198–208, Jan. 2018.

[6] 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.,* vol. 69, no. 12, pp. 4860 - 4871, Dec. 2

[7] J. S. P. Giraldo, S. Lauwereins, K. Badami, and M. Verhelst "Vocell: A 65-nm speech-triggered wake-up soc for 10 Î¼w keyword spotting and speaker verification," *IEEE J. of Solid-State Circuits,* vol. 55, no. 4, pp. 868-878, Apr. 2020.

[8] United States Environment Protection Agency, "Green-housegas equivalenciescalculator," [Online], https://www.epa.gov/energy/ Mar. 2022