A Real-Time Reconfigurable AI Processor Based on FPGA

Yue Ri Jeong, Kwonneung Cho, Youngwoo Jeong, Sun Beom Kwon and Seung Eun Lee*
Department of Electronic Engineering
Seoul National University of Science and Technology
Seoul, Republic of Korea
{jeongyueri, chokwonneung, jeongyoungwoo, kwonsunbeom*seung.lee}@seoultech.ac.kr

***Abstract*— As an AI application requires a lot of resources, optimization of hardware and software to target application is essential. Unlikely the software updates, it is hard to reconfigure the hardware architecture to the target application due to the static characteristics. In this paper, a real-time reconfigurable AI processor based on FPGA is proposed. The AI processor includes a reconfiguration block to update the FPGA and enables hardware reconfiguration with reasonable logic resources. The proposed reconfigurable AI processor was successfully implemented, and the feasibility was demonstrated by experimenting with the accuracy in various applications.**

Keywords*— *AI application, optimization, reconfigurable AI processor, hardware reconfiguration.

I. INTRODUCTION

Artificial Intelligence(AI) technology is applied to various applications and is widely employed in real life. As an AI utilizes a large amount of data and resources, high performance computing with application optimized hardware and software is essential [1]. A lot of research for optimizing the architecture of AI systems to specific applications understudying in order to achieve the design goals [2]. Unfortunately, the optimized hardware architecture and software algorithm depends on the target applications. Therefore, flexible reconfiguration and updates are required for AI systems when the target application is changed [3]. The software updates are relatively convenient compared to the hardware reconfiguration due to the flexible and reusable characteristics [4]. However, the hardware reconfigurations such as the architecture of AI processor or hardware resources are challenging and inconvenient due to the static characteristics. As the architecture of AI processor affects the performance and power budget of AI systems, hardware reconfigurable AI processor have advantages in terms of application specific optimization.

In this paper, a field programmable gate array (FPGA) based reconfigurable AI processor is proposed. The AI processor includes dedicated reconfiguration control modules and is able to be reconfigured in real-time with reasonable hardware costs. The reconfiguration of the AI processor is implemented with a microcontroller unit (MCU) and FPGA co-designed architecture, providing remote hardware updates. To demonstrate the benefits of hardware reconfiguration, we analyze the performance of the AI processor with image-based application and sound-based application according to the hardware resources of the AI processor. The analysis results demonstrate that the reconfigurable AI processor has advantages for optimization when the target application is changed.

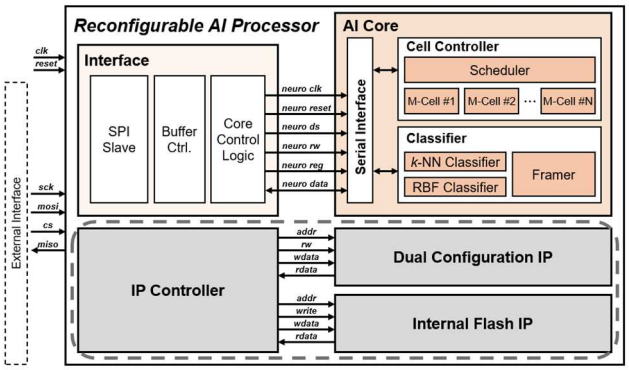


Fig. 1. Architecture of proposed AI processor

II. ARCHITECTURE

A. AI Processor

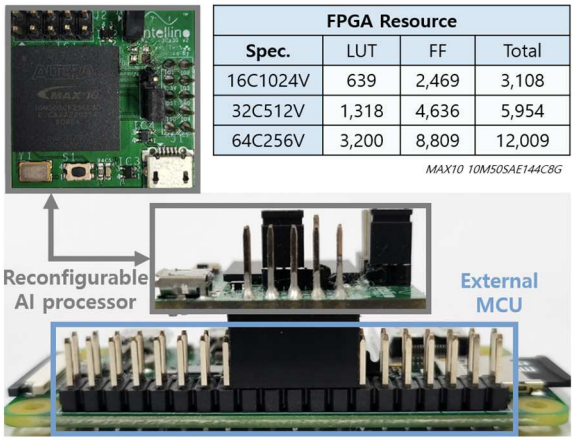
Fig.1 shows the overall architecture of the reconfigurable AI processor. The AI processor is consist of the interface, AI core, and reconfiguration block. The interface contains an SPI slave, buffer controller, and core control logic. The SPI slave receives data from the external MCU and sends the data to the buffer controller. The buffer controller stores the data received from the SPI slave to frame instruction. The core control logic decodes the instructions and generates control signals for learning and inference. The AI core performs learning and inference according to the control signal generated by the core control logic. In the AI core, the memory cells(M-cell) are employed to train and recognize data. Each cell includes a dedicated memory to store training data and a category value. In the inference process, M-cells calculate the distance between the trained data and recognition data in parallel, accelerating the distance calculation. Finally, the inference result is classified in the classifier module by comparing the distance values.

B. Reconfiguration block

The FPGA synthesizes the proposed AI processor by loading configuration data into the internal configuration RAM(CRAM). The configuration data are stored in configuration flash memory(CFM), and the FPGA loads the data into the CRAM in a power-on sequence or remote update state. For real-time reconfiguration, the AI processor includes the IP controller, dual-configuration IP, and internal flash IP. The

2023 IEEE International Conference on Consumer Electronics (ICCE) | 978-1-6654-9130-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICCE56470.2023.10043575

978-1-6654-9130-3/23/$31.00 ©2023 IEEE



FPGA Resource			
Spec.	LUT	FF	Total
16C1024V	639	2,469	3,108
32C512V	1,318	4,636	5,954
64C256V	3,200	8,809	12,009

MAX10 10M50SAE144C8G

Fig. 2. Experimental environment and results

internal flash IP provides a CFM access channel and informs the IP controller of the state of CFM. The dual-configuration IP triggers the remote update sequence to activate the CRAM. The IP controller is connected to the employed IPs with a bus interface and controls the IPs by accessing the IP registers.

The reconfiguration is performed in a series of sequences. First, the IP controller receives the configuration data from the external MCU according to the command decoded by the interface. The IP controller sets the internal flash IP and writes the configuration data to the CFM. In the CFM writing process, data error sparsely occurs. The IP controller accesses the same address and checks the validity of the written data to deal with the error. After storing all the configuration data, the IP controller sets the dual-configuration IP and triggers remote updates.

III. Experiment

As shown in Fig. 2, the experiment was performed with the FPGA and the external MCU. An Intel MAX10, 10M50DCF256C8G was utilized, and the proposed AI processor is synthesized on the FPGA. The MCU pre-processes the application data and transmits the data to the AI processor for learning and inference. In addition, the MCU sends the configuration data to the AI processor for reconfiguration. The AI processor was synthesized and experimented with various specifications. The FPGA resource table in Fig. 2 shows the total logic size according to each memory size and the number of the memory cells of the AI processor. The total logic size increases as the number of memory cell increases because the memory cell includes operators. The logic cells of reconfiguration blocks account for 62 LUTs and 131 FFs, which is a small amount of logic size compared to the AI processor.

The AI processor is able to employ various hardware resources according to the number of memory cells and the memory size of the cell. Therefore, the optimized architecture of the AI processor differs from the target application. Fig. 3 presents the accuracy of the AI processor in sound-based application and image-based application [5, 6]. The sound-based application recognizes the crashing window sound data, and the highest accuracy is 81.6% when employing 32 M-cells. However, the image-based application, which recognizes image

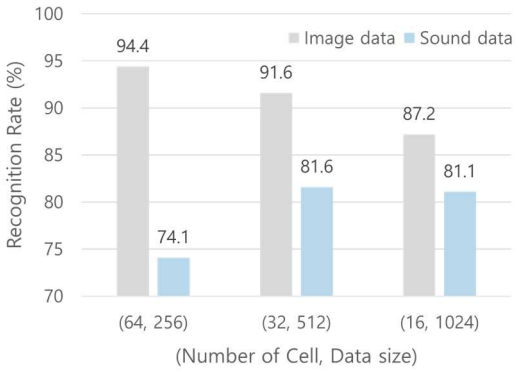


Fig. 3. Accuracy of each application

data of automobiles, shows the highest accuracy when 64 M-cells are employed. The analysis results demonstrate that the optimized specification of AI processor is different for each application. Therefore, hardware reconfiguration is efficient for various applications. Through the experiment, it was confirmed that the hardware reconfiguration was performed successfully according to the environment of the number of memory cells.

IV. Conclusion

In this paper, a FPGA-based hardware-reconfigurable AI processor is proposed. The reconfigurable AI processor changes the hardware architecture to employ various applications more efficiently. The reconfiguration is successfully implemented with optimized hardware resources. The validity of the proposal is demonstrated by experimenting with various specifications of AI processor in the sound-based application and image-based application. Therefore, the proposed AI processor enables reconfiguration with reasonable costs.

Acknowledgment

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program (IITP-2022-RS-2022-00156295) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation)

References

[1] Y. Ma, Y. Cao, S. Vrudhula and J. -s. Seo, "Optimizing the Convolution Operation to Accelerate Deep Neural Networks on FPGA," IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 26, no. 7, pp. 1354-1367, July 2018.

[2] Yoon, Y.H.; Hwang, D.H.; Yang, J.H.; Lee, S.E. "Intellino: Processor for Embedded Artificial Intelligence," Electronics 2020, 9, 1169.

[3] Yoon, Y.H; Oh, J.H; Kim, J.K; Ihm, H.B; Jeon, S.H; Kim, T.H; Lee, S.E. "Remote In-System Reconfiguration for Automotive Device," IEEE International Conference on Consumer Electronics (ICCE), Jan. 2019.

[4] Cho, K.; Kim, J.; Choi, D.Y.; Yoon, Y.H.; Oh, J.H.; Lee, S.E. "An FPGA-Based ECU for Remote Reconfiguration in Automotive Systems," Micromachines 2021, 12, 1309.

[5] Go, Kwang; Han, Chang; Cho, Kwon; Lee, Sang Muk. "Crime Prevention System: Crashing Window Sound Detection Using AI Processor," IEEE International Conference on Consumer Electronics (ICCE), 2021,1,1109.

[6] Cho, K.N; Oh, H.W.; Lee, S.E. "Vision-based Parking Occupation Detecting with Embedded AI Processor," IEEE International Conference on Consumer Electronics (ICCE), 2021,1, 1109