Parallel Implementation of CNN on Multi-FPGA Cluster

Yasuyu Fukushima, Kensuke Iizuka, Hideharu Amano

Dept. of Information and Computer Science, Keio University, Yokohama, Japan

Email: fic@am.ics.keio.ac.jp

Multi-access Edge Computing

MEC (Multi-access Edge Computing)

It provides an IT service environment and cloud computing functions for users at the edge of an access network that includes multiple access technologies.

Being standardized for 5G mobile networks

Examples：Factory control, Smart city traffic management, Security control, and Offloading a heavy task from an edge device　etc.

Demand for AI applications such as Deep Learning

Convolutional Neural Network

・Convolutional Neural Networks (CNNs) are known for achieving high recognition accuracy in image / voice recognition tasks and object detection.

Problem：**General-purpose processors** and **single FPGAs** have problems such as lack of resources and low throughput.

Solution：To implement a CNN inference accelerator on a **multi-FPGA system**, which can be expected to achieve a highly efficient implementation.

Our Work

- Multi-FPGA system is suitable for use as an MEC platform.
  - Low cost and low power consumption.
  - It can handle timing-critical jobs.
  - It can handle multiple requests from multiple edge devices at a base station.

Our Work

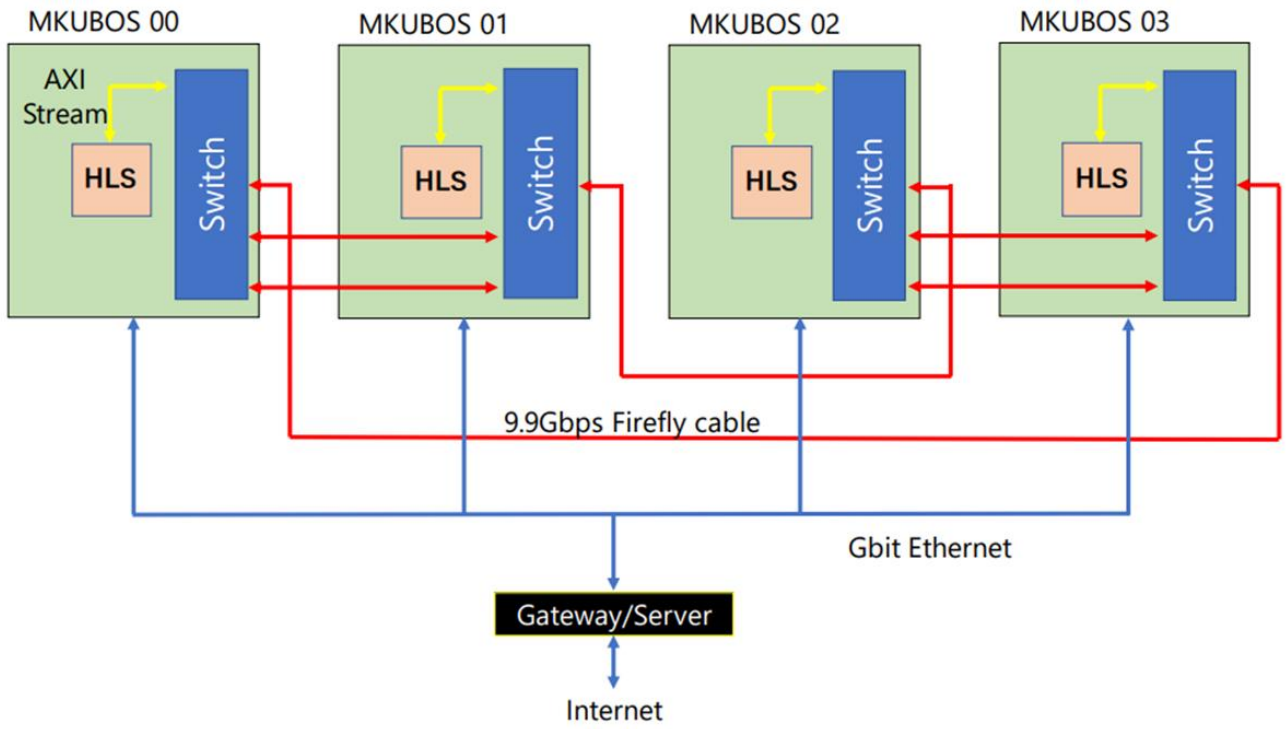
- As an example of image recognition for MEC application, ResNet-50, a typical CNN, was implemented on the multi-FPGA.

- The performance and power consumption of the actual system were evaluated.

Target Multi-FPGA System

M-KUBOS/PYNQ Cluster [1]

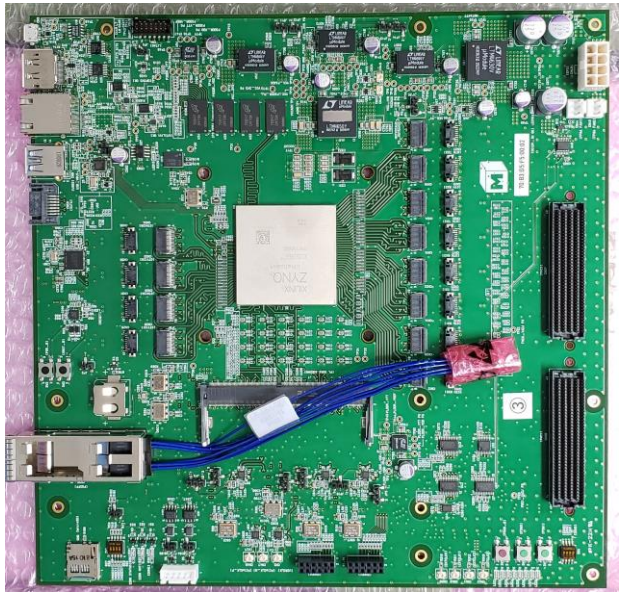
- We have developed an FPGA cluster by connecting PALTEK's four M-KUBOS boards with a high-performance, low-cost GTH serial link.
- An open source software platform called PYNQ has been introduced.

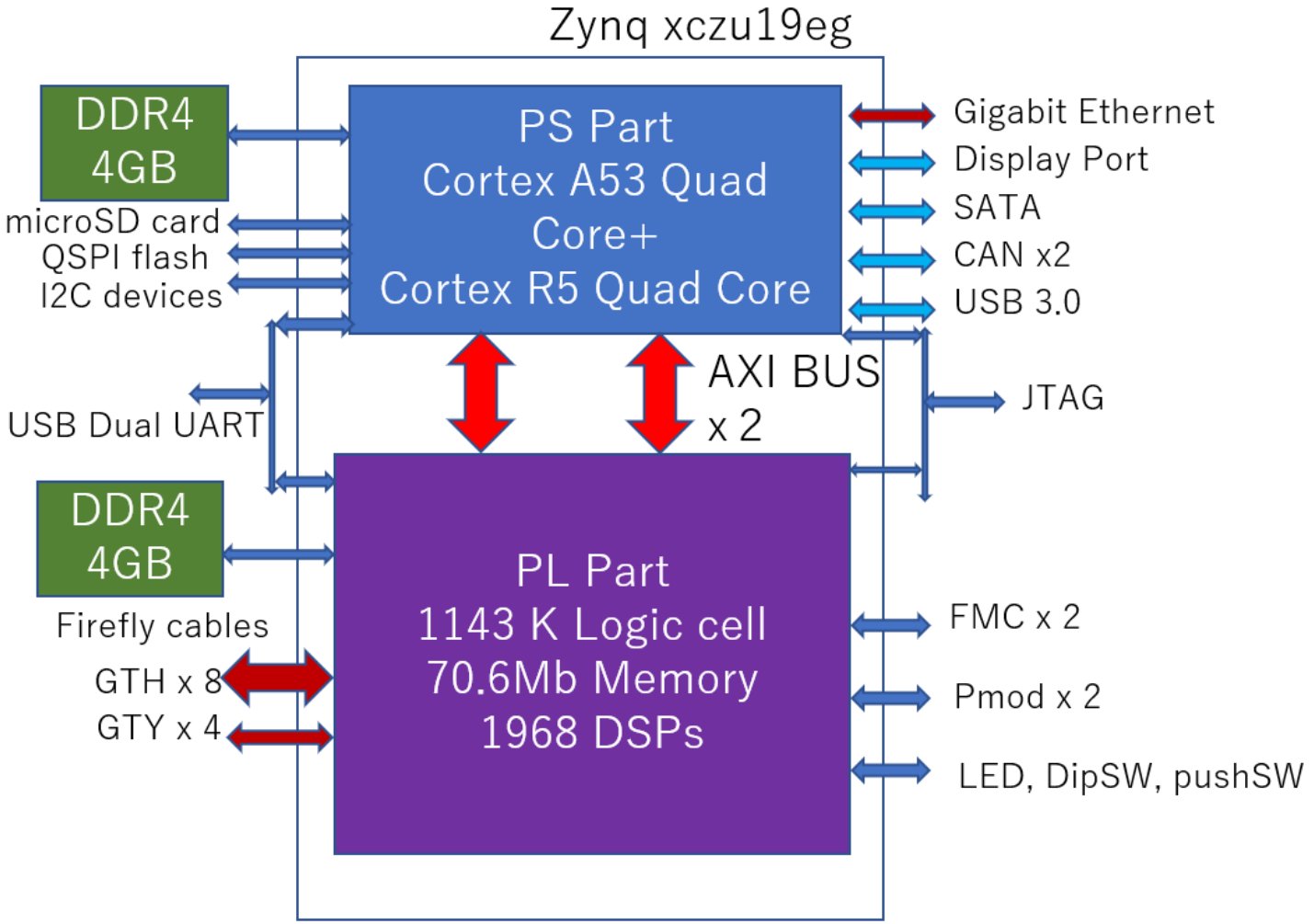


[1] Inage et.al. "MKUBOS/PYNQ cluster for Multi-access edge computing," CANDAR2021, Nov. 2021.

M-KUBOS

- Equipped with Zynq UltraScale+
- High-speed, two-way link with Firefly cable

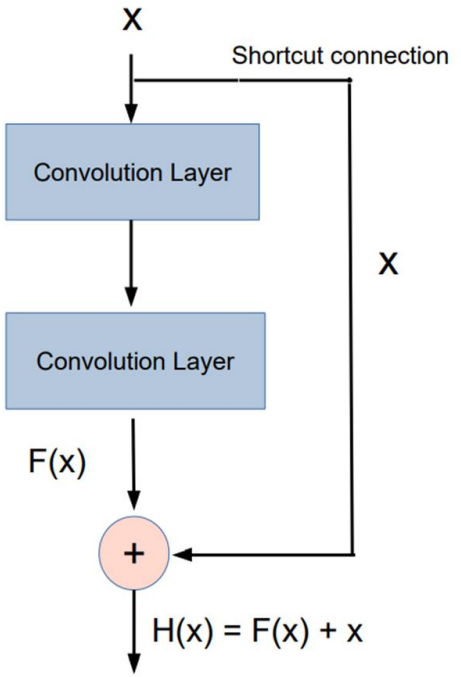




https://www.paltek.co.jp/mcube/index.html

ResNet-50

- ResNet [2] is one of the typical CNN models with a large amount of calculation and a large number of parameters.
- ResNet-50 was selected as a benchmark for MLPerf [3] and as a target for BrainWave [4].
- Operations：3.88×10^9
- Parameters：25.7 M

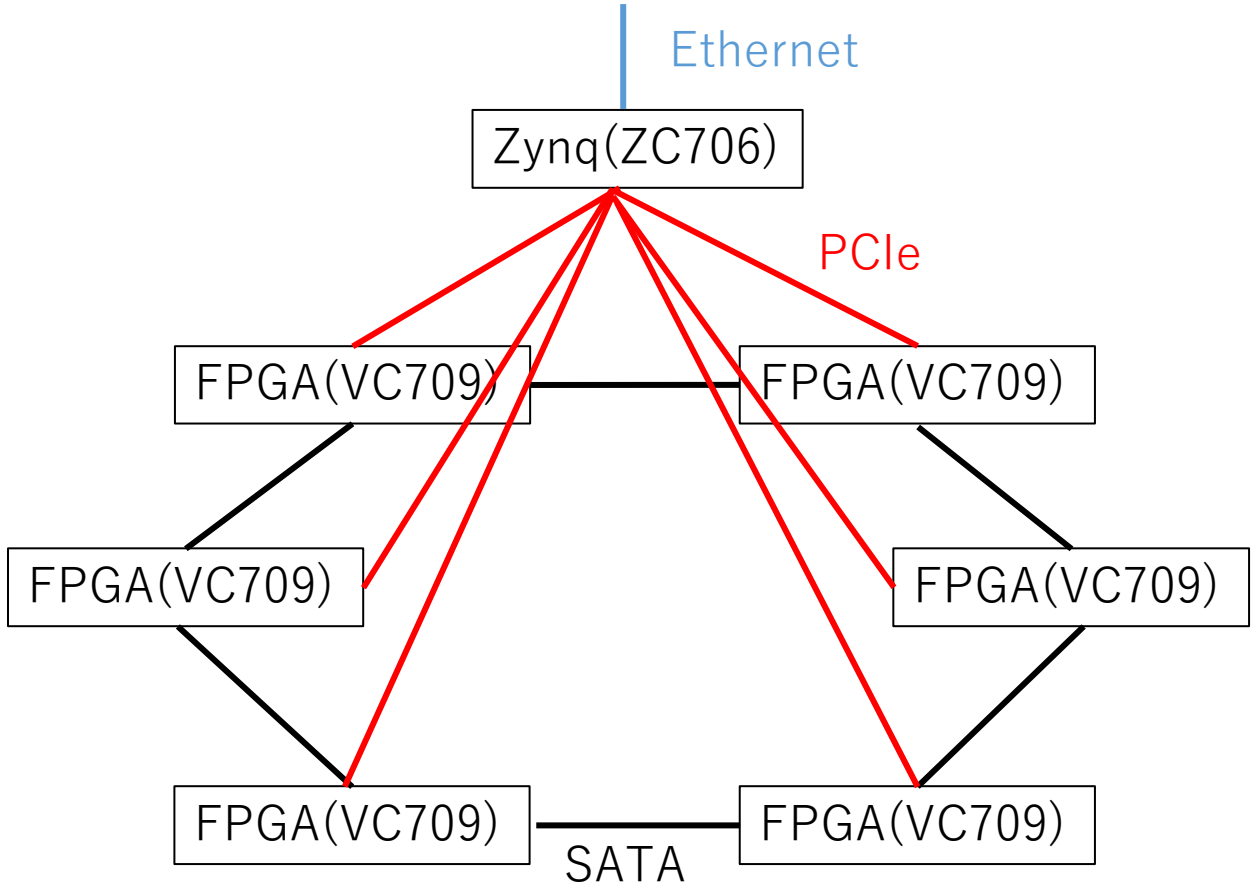


	conv1	max pool	conv2_x	conv3_x	conv4_x	conv5_x	average pool	1000-d fc
output size	112×112	56×56	56×56	28×28	14×14	7×7	1×1	1×1
block	$7 \times 7, 64$	$3 \times 3, 64$	$\begin{pmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{pmatrix} \times 3$	$\begin{pmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{pmatrix} \times 4$	$\begin{pmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{pmatrix} \times 6$	$\begin{pmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{pmatrix} \times 3$	-	-

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016, pp. 770–778.
[3] MLCommons, "MLPerf," https://mlcommons.org/ja/ (accessed 2021-6- 18).
[4] Microsoft Research, "Project Brainwave," https://www.microsoft.com/en-us/research/project/project-brainwave/ (accessed 2021-6-18).

Related Works [5]

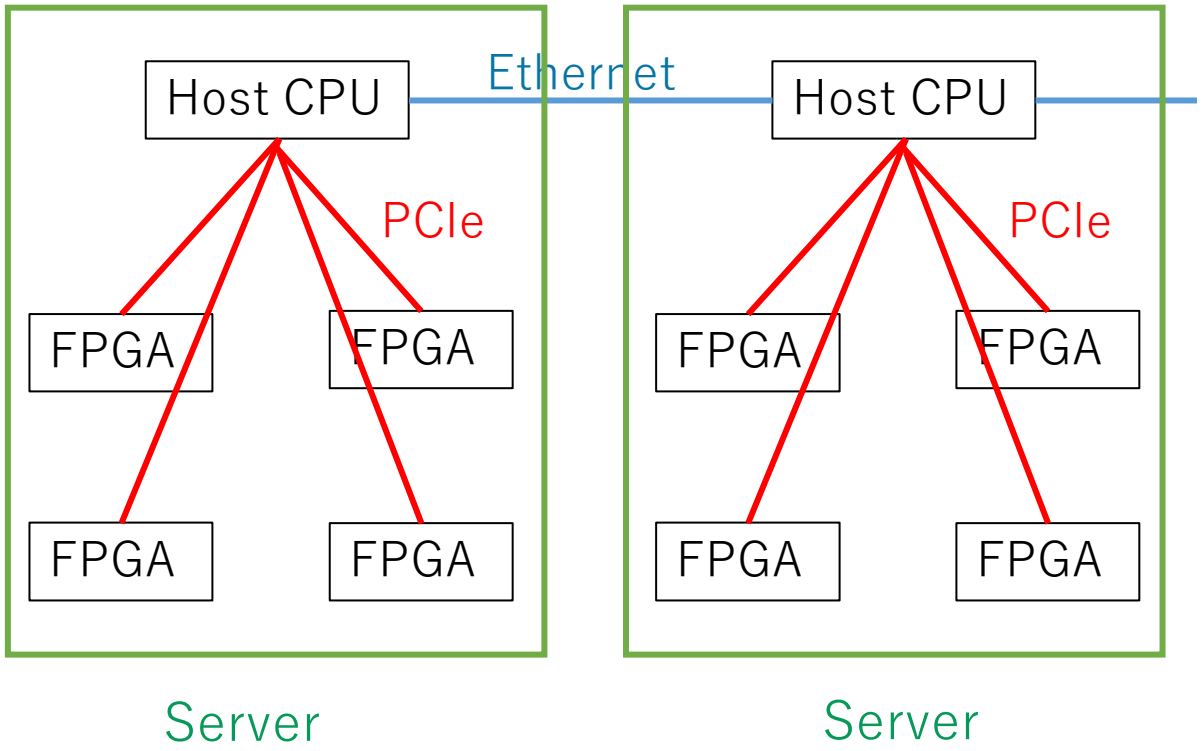
- Multi-FPGA system with one host Zynq board and six FPGAs connected
- Zynq and each FPGA are connected via PCIe
- Each FPGA is connected to two FPGAs via SATA
- The CNN models used in the evaluation are VGG-16 and AlexNet



[5] C. Zhang, D. Wu, J. Sun, G. Sun, G. Luo, and J. Cong, "EnergyEfficient CNN Implementation on a Deeply Pipelined FPGA Cluster," in Proceedings of the 2016 International Symposium on Low Power Electronics and Design, ser. ISLPED '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 326–331. [Online]. Available: https://doi.org/10.1145/2934583.2934644

Related Works [6]

- A multi-FPGA system that configures a server by connecting a host CPU and four FPGAs.
- Communication between servers is done via Ethernet.
- Host CPU and each FPGA are connected by PCIe.
- The CNN model used for evaluation is ResNet-152.



[6] W. Zhang, J. Zhang, M. Shen, G. Luo, and N. Xiao, "An Efficient Mapping Approach to Large-Scale DNNs on Multi-FPGA Architectures," in 2019 Design, Automation Test in Europe Conference Exhibition (DATE), March 2019, pp. 1241–1244.

Merit of PYNQ Cluster

Related work [5] ：FPGAs were connected by SATA.

➡ Communication between FPGAs was not fast.

Related work [6] ：FPGAs were not directly connected.

➡ Communication between FPGAs had to be managed by the host CPU.

PYNQ Cluster：FPGAs are directly connected by high-speed serial links.

➡ It enables stream processing without network bottlenecks and facilitates pipeline processing between boards.

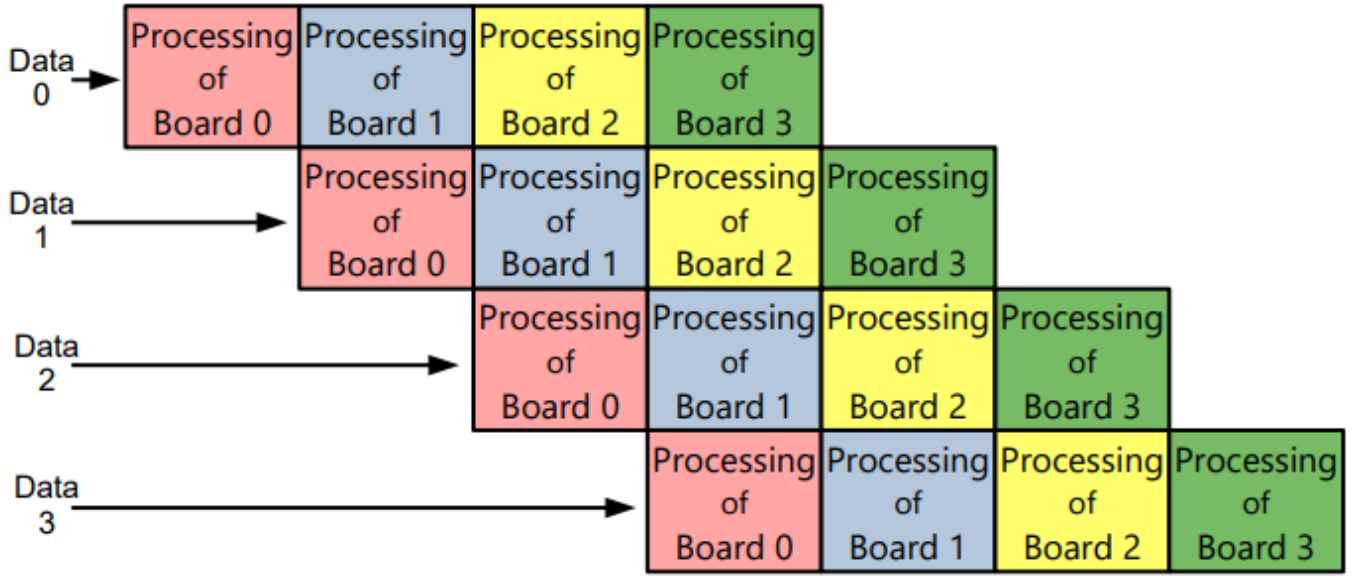
Implementing ResNet-50 on PYNQ Cluster

Approach

Allocate first convolution layer and maxpooling layer to board 0.

Allocate average pooling layer and FC layer to board 3.

Allocate the common design convolution layers to board 0-3.



To make the processing time of each board as equal as possible…

- Determine where we divide the network from the estimate by Vivado HLS.
- Fine-tune the division based on the execution time on the actual machine.

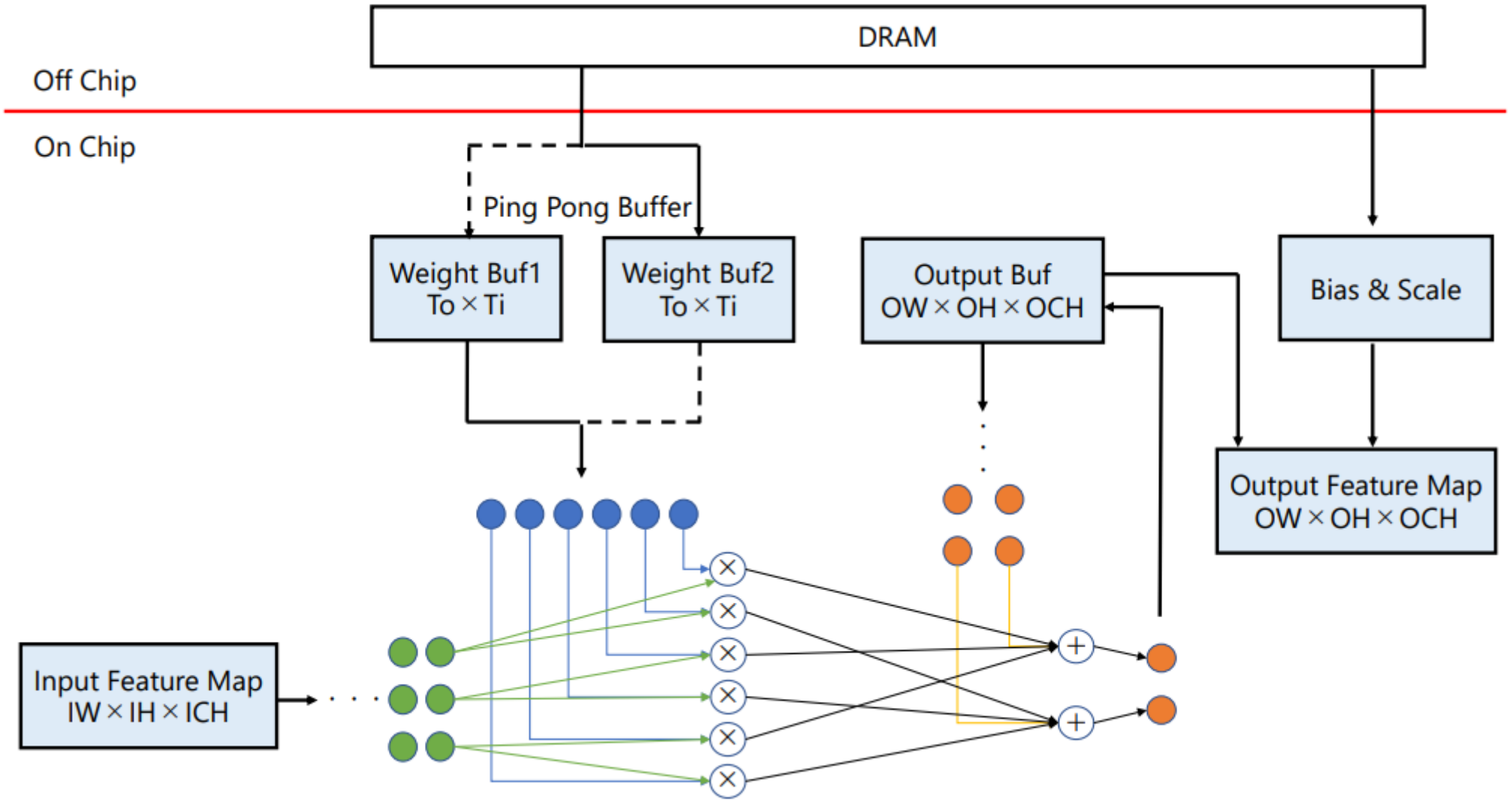
Quantization

- The model trained with a 32-bit floating-point number was quantized to an 8-bit integer number for both the weight and feature map.
- Quantization can reduce the requirements for memory bandwidth and on-chip memory resources.
- By using an 8-bit integer for both the weight and feature maps, two MAC operations can be performed with one DSP [7].

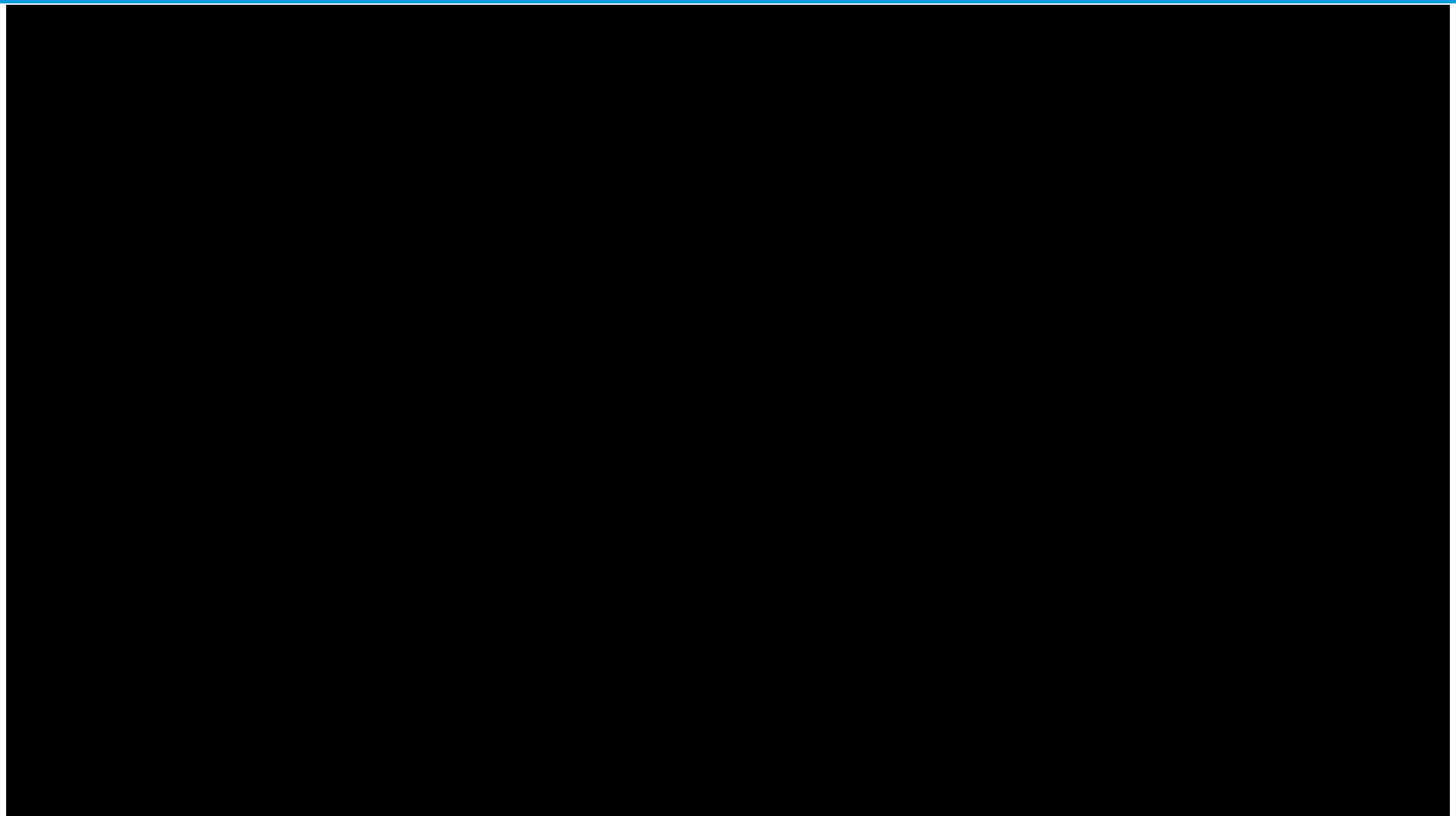
	Float32	INT8
Top-1 Accuracy	74.4%	74.2% (-0.2%)
Top-5 Accuracy	93.0%	92.8% (-0.2%)

[7] D. Nguyen, D. Kim, and J. Lee, "Double MAC: Doubling the performance of convolutional neural networks on modern FPGAs," in Design, Automation Test in Europe Conference Exhibition (DATE), 2017, 2017, pp. 890–893.

Architecture of CONV Layer

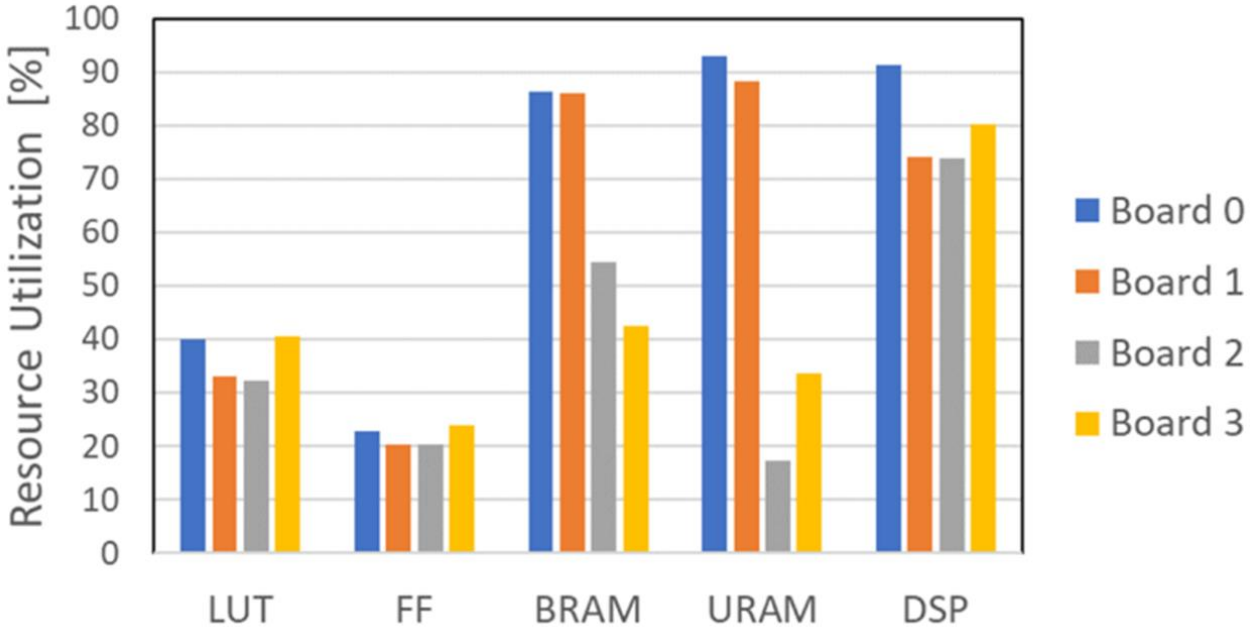


Demonstration



Experimental Result

Tools：Vivado HLS 2019.1.3, Vivado 2019.1.3



- The performance is determined by the part with the longest execution time in pipeline processing.

  ➡ It is possible to process one data every 13.3 ms.

- This implementation achieved throughput: 75.1 FPS, performance: 292 GOPS, and power efficiency: 5.15 GOPS / W.

	Board 0	Board 1	Board 2	Board 3
Execution Time (msec)	12.7	13.3	11.6	9.7
Power Consumption (W)	15.6	13.9	13.2	14.0

Performance Comparison

- This implementation achieved 17 times faster speed and 86 times more power efficiency than the **CPU** deep learning framework implementation.
- This implementation achieved 3.8 times more power efficiency than **GPU** deep learning framework implementation.
- Compared to the 4-board implementation of related work [6], this implementation achieved 4.6 times its performance.

	This Implementation	[6] (ResNet-152)	CPU	GPU
Device	Xilinx Zynq UltraScale+ (×4)	Xilinx Vertex UltraScale (×4)	AMD Ryzen Threadripper 3990X	NVIDIA GeForce RTX 3090
Frequency (MHz)	100	150	2900	1400
Precision	INT8	Fixed16	Float32	Float32
Performance (GOPS)	292	62.9	17	473
Power Efficiency (GOPS/W)	5.15	-	0.06	1.35

Conclusion

- As an example of image recognition for MEC applications, ResNet-50 was divided and implemented on a PYNQ cluster consisting of four M-KUBOS boards.
- The throughput was improved by performing pipeline processing between the four boards.
- The proposed implementation achieved a throughput of 75.1 FPS, performance of 292 GOPS, and power efficiency of 5.15 GOPS / W.