2022 IEEE 30th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM) | 978-1-6654-8332-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/FCCM53951.2022.9786215

Dual-Line-Systolic Array for High Performance CNN Accelerator

Peng Xue
Shenzhen Institute
of Advanced Technology,
Chinese Academy of Sciences
peng.xue@siat.ac.cn

Lunshuai Pan
Southeast University
ls.pan@siat.ac.cn

Litao Sun
Southeast University
slt@seu.edu.cn

Mingqiang Huang
Shenzhen Institute
of Advanced Technology,
Chinese Academy of Sciences
mq.huang2@siat.ac.cn

***Abstract*—Systolic array has been the crucial architecture for accelerating convolutional neural networks (CNN) since the success of Google's TPU (Tensor Processing Unit). In this work, we propose high throughput and low delay dual-line-systolic array for accelerating the convolutional neural networks. With the line-by-line vector-style systolic dataflow, the peripheral circuit can be well simplified and the loading/offloading delay can be greatly reduced. Besides, to fully take advantage of the DSP (Digital signal processor) INT8 computation in FPGA, dual-line-systolic array is developed, by which the computation throughput can be doubled. Finally, the proposed accelerator is deployed on PYNQ-Z2 for practically accelerating VGG16 neural network, peek throughput of the convolution layer can reach as high as 107.21 GOPS, which has exceeded all of the previous works on the same hardware platform.**

I. Analysis of Dual-Line-Systolic Array

Deep convolutional neural networks (CNN) have been widely adopted to address many artificial intelligence tasks successfully [1].We propose the line-systolic array (Fig. 1b), where the input-feature data delivers vertically in the PE array in a row-by-row vector style, the weight data is first horizontally loaded and then fixed at each PE, and the partial-sum data is perpendicularly processed to its output. Furthermore, taking the insufficient resources of edge devices into account, we maximize the utilization of existing resources through DSP multiplexing and thus create the dual-line-systolic array(Fig. 1c). The difference is that the dual-linesystolic transfers two rows of weight at each cycle. Xilinx's DSP architecture is optimized for INT8 deep learning inference and each DSP48E2 slice can pack two parallel INT8 multiply-add operations.

Fig. 1d takes the following multiplication as an example: $(A \times 2^{17} + D) \times B$, where A and D are 8-bit signed weights and B is 8-bit signed feature. The lowest 16-bit data of this product is the value of $A \times B$, and the highest 16-bit data is the value of $D \times B$. In summary, the proposed dual-line-systolic will take full advantage of DSP and achieve a good balance between fan-out and delay.

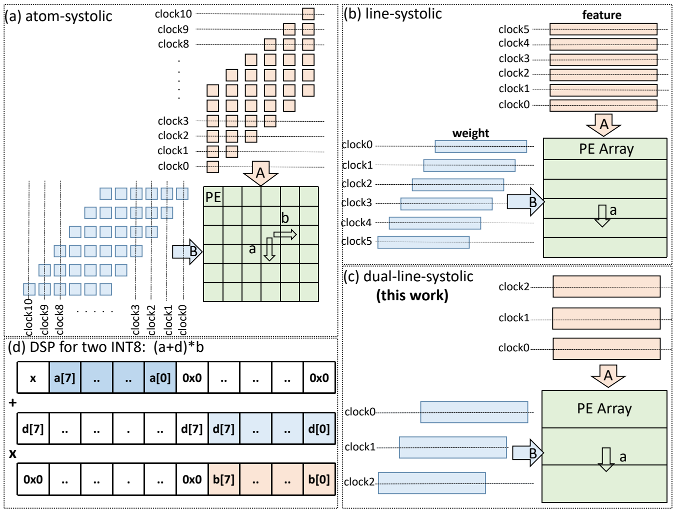

Fig. 1. Data flow in different systolic architectures: (a) traditional atom-systolic, (b) dual-line-systolic (this work) and (d) DSP multiplexing.

II. Results

Before the practical test on development board, three kinds of systolic array are synthesized using Vivado (version 2018.1). Considering that the PYNQ-Z2 contains only 220 DSP on board, We synthesized the 256-parallelism firstly and all three modes can successfully generate the bitstream with some extra LUTs. Once the parallelism was increased to be 512 (T_{in}×T_{out}=32×16), only the purposed dual-line-systolic array can be successfully implemented, while the others are all failed due to the limited resources in FPGA.

We conduct the test on PYNQ-Z2 board to analyze the practical throughput under 120MHz working frequency. Since the computation parallelism is 512, the theoretical MAC operation performance of our accelerator will be deduced by 120MHz×2×512ops = 120 Gops. The practical peek throughput is 107.21 GOPS which is close to the theoretical limit, and the average throughput of all convolution layers is 65.12 GOPS. Compared with atom-systolic and line-systolic, the practical peek throughput of dual-line systolic array is 1.9 times higher and the average throughput is 1.53 times higher.

References

[1] M. Dhouibi, A. K. Ben Salem, and S. B. Saoud, "Cnn for object recognition implementation on fpga using pynq framework," in *2020 IEEE Eighth International Conference on Communications and Networking (ComNet)*, 2020, pp. 1–6.

978-1-6654-8332-2/22/$31.00 ©2022 IEEE