BF16-to-BFP Conversion Architecture for Systolic Array Computation

Jihyeon Hwang, Taehong Min, Jungwoo Park and Seung Eun Lee*

Dept. of Electronic Engineering, Seoul National University of Science and Technology, Seoul, Republic of Korea

Various AI accelerators have been developed to improve computational efficiency, and the systolic array has been widely adopted for its massive parallelism and data reuse. Among the commonly used numerical formats, the brain floating point 16-bit (BF16) provides high precision. However, BF16 still suffers from hardware complexity and high power consumption, making it less suitable for systolic array operations [1].

To overcome these limitations, we propose the block floating point (BFP) converter that maintains the accuracy of BF16-based data while improving computational efficiency. The converter determines a shared exponent and aligns mantissas to generate BFP data, simplifying operations. It supports various formats by configuring widths. In this process, the mantissas are shifted based on the maximum exponent, and rounding is performed using the discarded bits. If an overflow occurs after rounding, re-alignment is performed. The converted data are then processed by a 4×4 systolic array, which performs only mantissa computations to reduce complexity while maintaining accuracy. The overall architecture is illustrated in Fig. 1.

The proposed architecture was designed in Verilog HDL and implemented on an FPGA using the ZCU104 development board. We evaluated performance by comparing the accelerator with the conventional BF16-based operation. The proposed architecture achieved approximately 2.17× lower resource utilization and 2.07× higher computation speed than the BF16-based operation. The comparison results are summarized in Table I. These results demonstrate that the conversion architecture overcomes the limitations of the BF16 format by employing BFP format conversion, thereby optimizing the systolic array computation process.

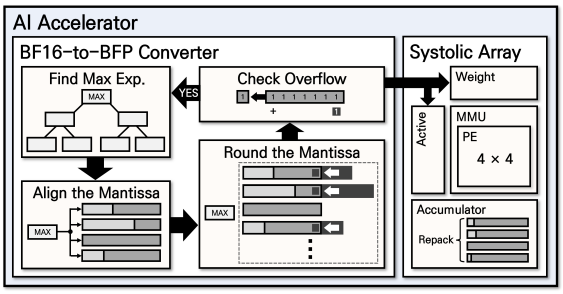

Fig. 1. The overview of the proposed architecture

Table I. Resource utilization and computation speed

Resource	BF16-based operation	Ours
LUT	17,064	7,129
Register	2,286	1,797
Total Resources	19,350	8,926
Comparative Utilization	**2.17×**	**1×**
f_{max} [MHz]	24.39	50.55
DSP	16	16
Comparative Speed	**1×**	**2.07×**

Acknowledgments This work was supported by the IITP(Institute of Information & Coummunications Technology Planning & Evaluation)-ITRC(Information Technology Research Center) grant funded by the Korea government(Ministry of Science and ICT)(IITP-2025-RS-2022-00156295)

References [1] S. Q. Zhang, B. et.al "FAST: DNN Training Under Variable Precision Block Floating Point with Stochastic Rounding," in Proc. 28th IEEE Int. Symp. on High-Performance Computer Architecture (HPCA-28), Feb. 2022.