BitSys: Bitwise Systolic Array Architecture for Multi-precision Quantized Hardware Accelerators

Yuhao Liu[1,3], *Student Member, IEEE*, Salim Ullah[2], Akash Kumar[2,3], *Senior Member, IEEE*
[1]Dresden University of Technology, Germany [2]Ruhr University Bochum, Germany
[3]Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI Dresden/Leipzig), Germany
Email: {yuhao.liu1, salim.ullah}@tu-dresden.de, akash.kumar@rub.de

I. Extended Abstract

Quantized Neural Networks (QNN) have been widely applied in hardware accelerator designs for edge. Because lower precision in quantization leads to higher accuracy loss, the mixed-precision scheme has been explored by using different precision in different layers to trade off resource consumption and inference accuracy. Because regular multiplier designs do not support the reconfiguration for multi-precision, we explored a runtime reconfigurable multi-precision bitwise systolic array design, *BitSys*, for mixed-precision multiplication in QNN accelerators. The popular design in previous works, such as [1], divides the inputs of multipliers as two parts for four sub-multipliers and preset left shifting, achieving the reconfigurable multiplication by disabling two of the sub-multipliers. Our design is inspired by the *Bitshifter* architecture from the works of Liu et al. [2, 1]. We convert the $n \times n$-bit multiplication as $A \times B = \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} 2^{i+j} a_i b_j$. As shown in Figure 1, partial products P_{i+j} is the sum of sub-partial products $a_i b_j$ with left shifting value, $i + j$. The sub-partial product masks select the desired $a_i b_j$ to configure the multi-channel according to the precision. One mask square represents one channel. Therefore, we can implement a bitwise systolic array as Figure 2 (left). The sub-partial product computation and mask are fused in one LUT primitive as one processing element in Figure 2 (right up). The sums of elements, considered the sign-bits, in the diagonal with the same left shifting value shown in Figure 2 (left) are the inputs, D_k, of the output-generate pipeline of Figure 2 (right down). Systolic array sequentially generates the D_k, and the final multi-precision output is the sum of all D_k. We implemented our *BitSys* multiplier as a systolic array for mixed-precision QNN acceleration. The comparison with the works of Liu et al. [1] is shown in Table I. Our systolic array accelerator consumes more hardware resources than the three single-layer accelerator instances of Liu et al. [1]. However, because of our bitwise processing design, *BitSys* instance can support $250MHz$ clock frequency because of the low critical path delay of our multiplier and achieves 188.5-274.7% speed-up in the evaluation of one four-layer 1/2/4/8-bit mixed-precision quantized MLP. Furthermore, our design does not change the input/out width when configuring to different precision, which can be easily integrated into existing accelerator designs.

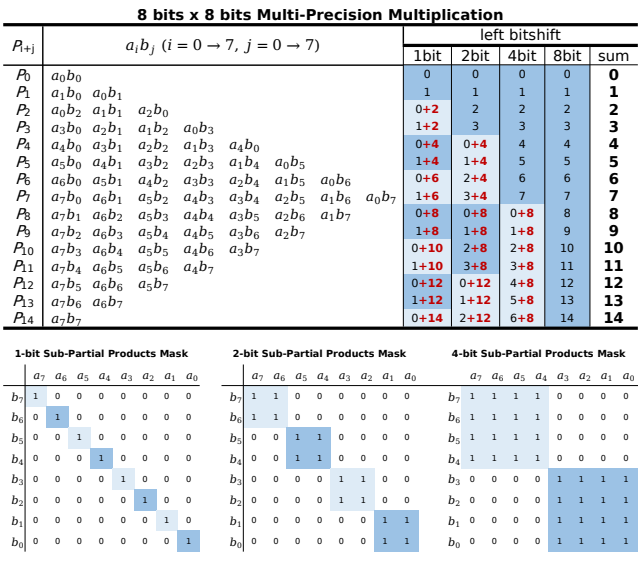

Fig. 1: Tree Structure of Multi-precision Accumulator

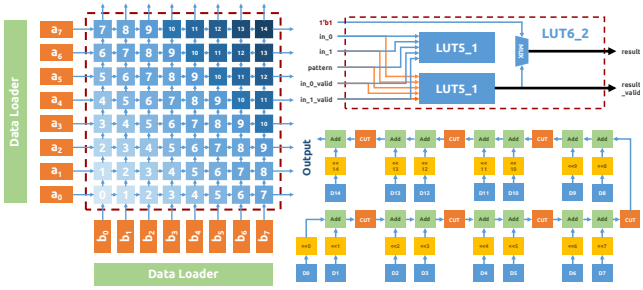

Fig. 2: Tree Structure of Multi-precision Accumulator

TABLE I: Comparison of Resource Consumption and Latency

Design	Precision	LUT	FF	BRAM	Frequency	Latency/μs
Vivado IP [1]	8/8/8/8 1/2/4/8	24090	22175	135	$150MHz$	137.654 131.059
Bitshifiter [1]	1/2/4/8	42952	22486	138	$125MHz$	56.658
Multiplier Tree [1]	1/2/4/8	37020	22500	138	$100MHz$	69.27
BitSys	1/2/4/8	44468	64176	139.5	$250MHz$	36.741

References

[1] Yuhao Liu et al. "High Flexibility Designs of Quantized Runtime Reconfigurable Multi-Precision Multipliers". In: *IEEE Embedded Systems Letters* (2023), pp. 1–1.

[2] Negar Neda et al. "Multi-Precision Deep Neural Network Acceleration on FPGAs". In: *2022 27th Asia and South Pacific Design Automation Conference (ASP-DAC)*. IEEE. 2022, pp. 454–459.