Reconfigurable AI Accelerator with 2D Systolic Array

MAC Group: Jacob Brown, Yuqi Lan, Kunal Bhandarkar

Motivation

This project aims to design a reconfigurable 2D systolic array based AI accelerator to achieve efficient, scalable, and performance AI inference on constrained hardware.

2D Systolic Array

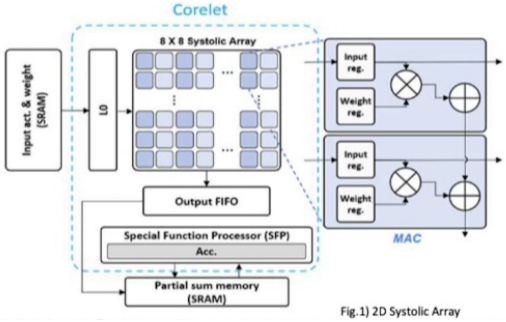


Fig.1) 2D Systolic Array

VGGNet and ResNet

The table below details the two models trained via quantize-aware training and evaluated on the CIFAR-10 dataset

	VGG16	ResNet20
Accuracy (CIFAR 10)	9019/10000 (90%)	9058/10000 (91%)
Quantization Error	3.2451487186335726e-07	0.00013859561295248568

Mapping on FPGA (Cyclone IV GX)

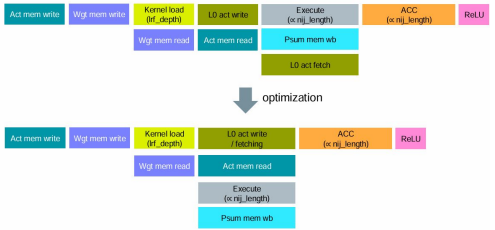
	Weight Stationary Array	Reconfigurable Array
Fmax	125.88 MHz	115.55 MHz
Switching Activity	20%	20%
PVT	1200 mV 100C	1200 mV 100C
Total Thermal Power	229.58 mW	234.83 mW
Total Logic Elements	22,647	28,018
Total Registers	12,266	14,619
Total Operations	128 operations	128 operations
GOPs/s	16.11 G operations per second	14.79 G operations per second
TOPs/W	557.53 operations per Watt	545.07 operations per Watt

Alpha 1: ResNet20

We trained a 80% sparsity ResNet20 model and split it into 2x2 tiles. By testing this dataset on 2D Systolic Array we can verify the efficiency of pruning and the correctness of tiled convolution.

Alpha 2: FIFO Parallelism

We can write and read L0 when executing. We can also read OFIFO when it is ready. These optimizations reduce cycle count and reduce both FIFO depths to 16 (4X smaller).



Alpha 3: Accumulation during OFIFO read

We can start to perform our accumulation while reading the psums from the OFIFO and save them with dual-port SRAM to cut down on cycle count.

Our Corelet →