CNNs on Multi-channels 2-D Systolic Array with Versatile Pruning

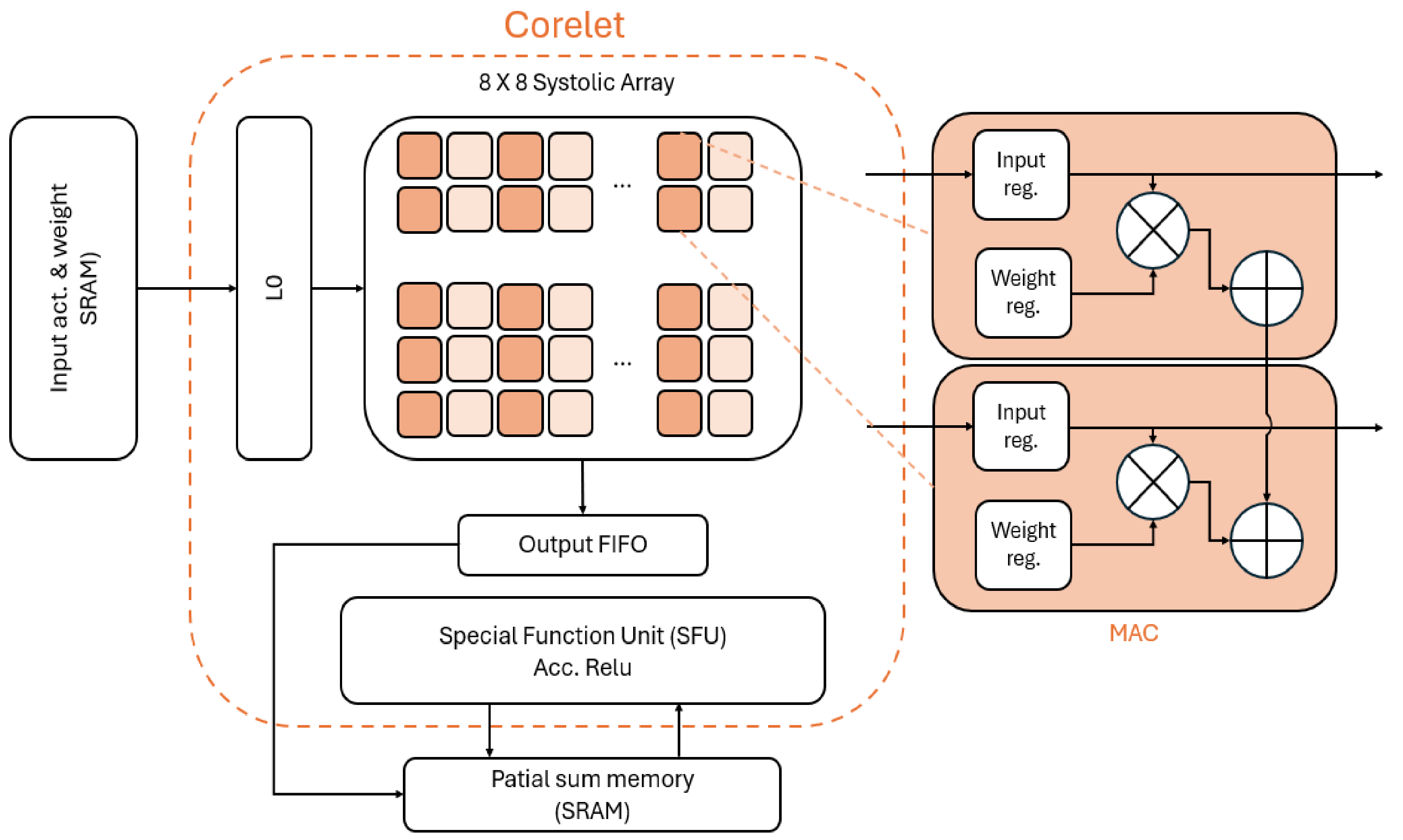
Starlight Team

Jingbin Lin, Haotian Ye, Liangyuan Wang, Zhuoting Yang, Zihao Yang

Motivation

With their high parallelism and flexibility, 2-D systolic arrays are an attractive platform for low-precision AI acceleration. In this project, we build a reconfigurable 8×8 2-D systolic array with 2/4-bit support and WS/OS dataflows, deploy it on a Cyclone IV GX FPGA to accelerate quantized VGG-style networks on CIFAR-10.

2-D Systolic Array



VGGNet with quantization-aware training

Model / Layer	Value
Dataset	CIFAR-10
4-bit Act Model Acc	91.2%
2-bit Act Model Acc	90%
4-bit Act Quantization Error	3.2444e-07
2-bit Act Quantization Error	1.9340e-06

Software

Alpha 1. Optimized training for Quantized VGG

Method	4-bit Acc	2-bit Acc
Baseline(SGD +Momentum)	89%	89%
Adam + Label Smoothing + Cosine Scheduler	91.2%	90%

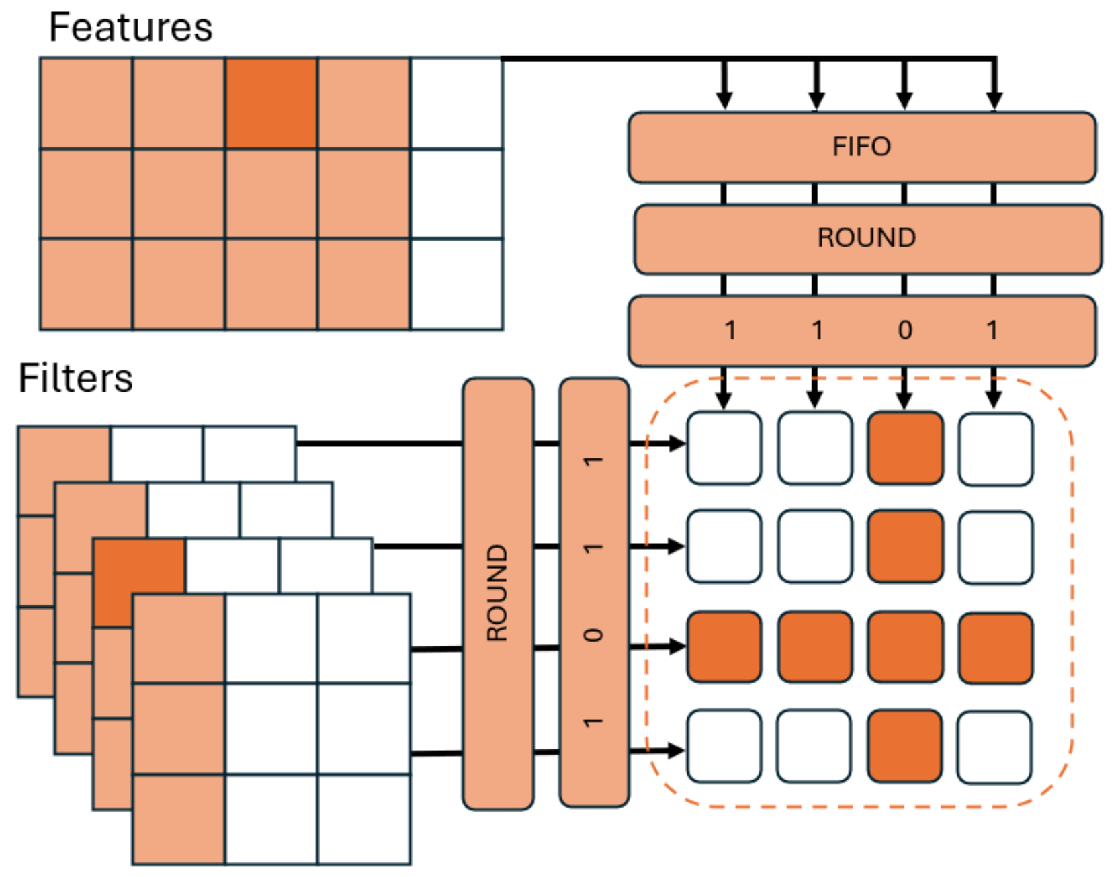
Alpha 2. The C2F Pruning Method (Mixed-Granularity)

Reach unstructured pruning sparsity, keep high precision while keep systolic array computation friendly as structured pruning.

Model Precision	VGG16_Quanty (4bit)	Resnet20_Quant(4bit)
Unstructured Pruning	90.31%	89.94%
Structured Pruning	83.48%	78.53%
Our Method (Coarse-to-Fine)	90.22%	87.74%

Model Sparsity	VGG16_Quanty (4bit)	Resnet20_Quant(4bit)
Unstructured Pruning	87.90%	69.82%
Structured Pruning	96.24%	61.92%
Our Method (Coarse-to-Fine)	87.92%	81.64%

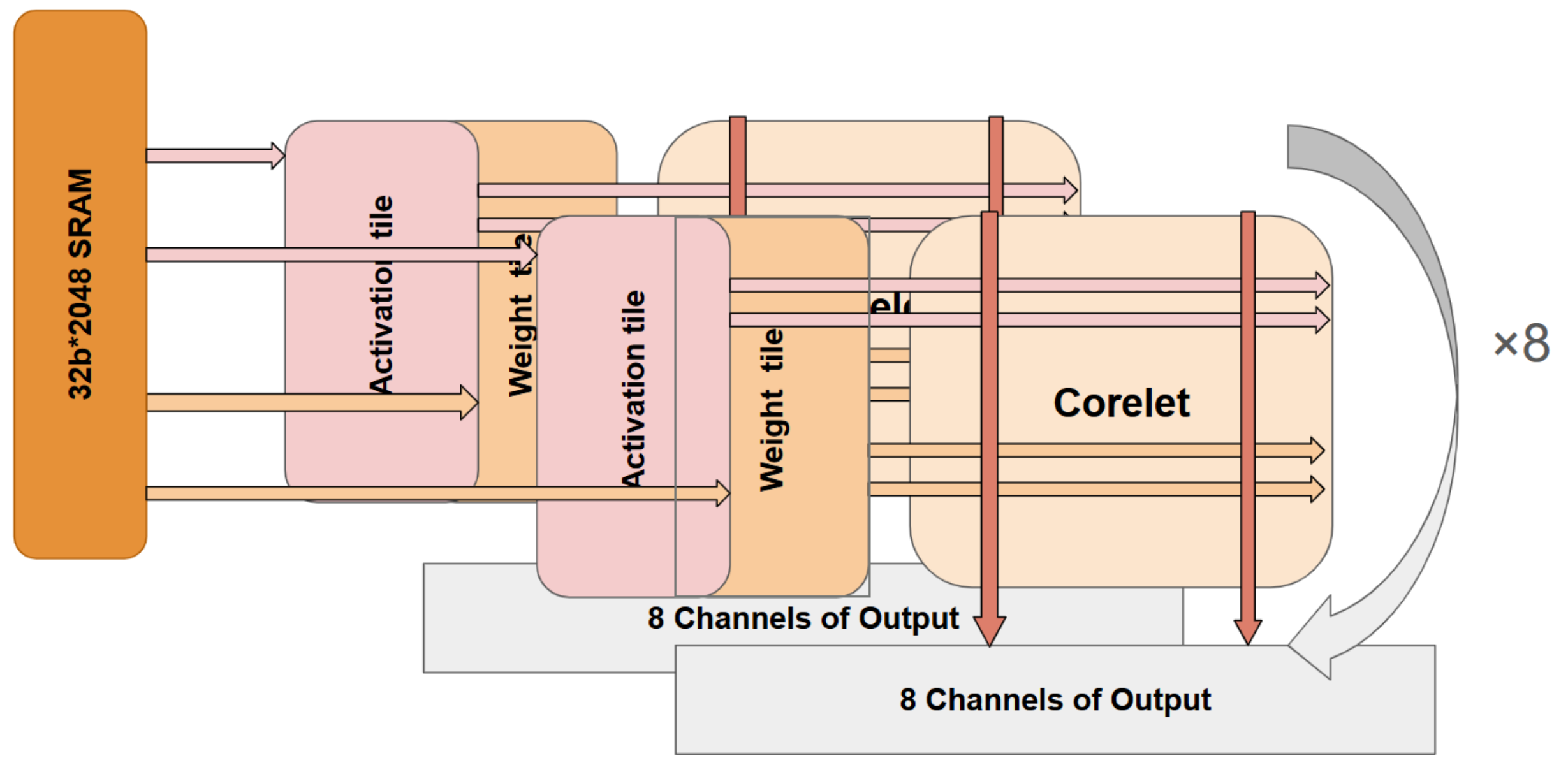
Alpha 3. Output Stationary Skip Optimization(WIP)



We introduce fine-grained clock gating to disable the clock of PEs, FIFOs, and MAC datapaths during idle or skip cycles. This reduces unnecessary switching activity and improves power efficiency without affecting performance or timing. (in progress)

Alpha 4. Scalable Multi-cores Tiling (WIP)

We scale the 8×8 systolic array into multiple parallel cores that process tiled feature-map regions independently. A simple tiling controller distributes inputs and psums across cores with low overhead, enabling higher throughput and near-linear performance scaling while remaining compatible with WS/OS modes. (in progress)



Mapping on FPGA (Cyclone IV GX)

	VGG16
Ops	128
Frequency	128.72MHz
Dynamic power	340.72mW
GOPs / s	16.5
GOPs / W	48.4
Logic Elements	17,112 / 149,760 (11 %)

References

[1] C. Ogbogu et al., "Energy-Efficient ReRAM-Based ML Training via Mixed Pruning and Reconfigurable ADC," 2023 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED), Vienna, Austria, 2023, pp. 1-6