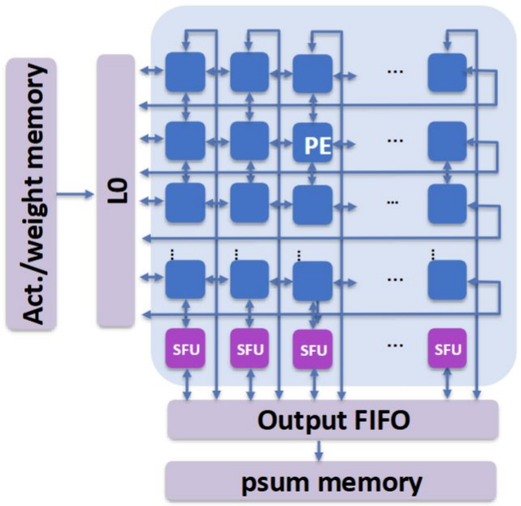
2D Systolic Array

Tensor Team: Alexis Yu, Jinshi He, Joey Liu, Justin Sin

Motivation:

AI applications demand efficient accelerators for high performance and energy efficiency. Systolic arrays, with their parallelism, are ideal for such tasks but require optimized designs for FPGAs like Cyclone IV GX. This project introduces a weight- and output-stationary systolic array to enhance throughput and energy efficiency for AI workloads.

2D Systolic Array:



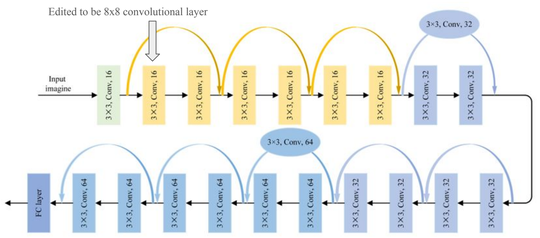
VGG16 with quantization-aware training:

	VGG16
Accuracy	91%
Quantization Error	3.9279e-7

Mapping on FPGA:

	VGG16
Frequency	132.21 MHz
Dynamic Power	32.42 mW
GOPs /s	1.003
GOPs /W	3.054
Logic Elements	16,595

Alpha 1: Incorporating ResNet20 Model



	ResNet20
Accuracy	80%
Quantization Error	5.0165

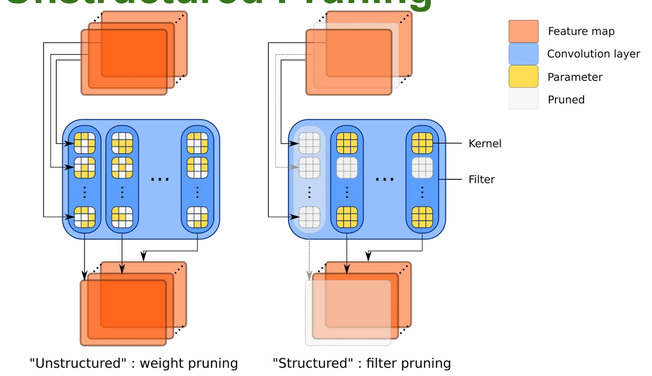
References

ResNet-20 Architecture. | *Download Scientific Diagram*, www.researchgate.net/figure/ResNet-20-architecture_fig3_351046093. Accessed 5 Dec. 2024.

Dash, Abhipraya Kumar. “VGG16 Architecture.” *OpenGenus IQ: Learn Algorithms, DL, System Design*, OpenGenus IQ: Learn Algorithms, DL, System Design, 19 Nov. 2020, iq.opengenus.org/vgg16/.

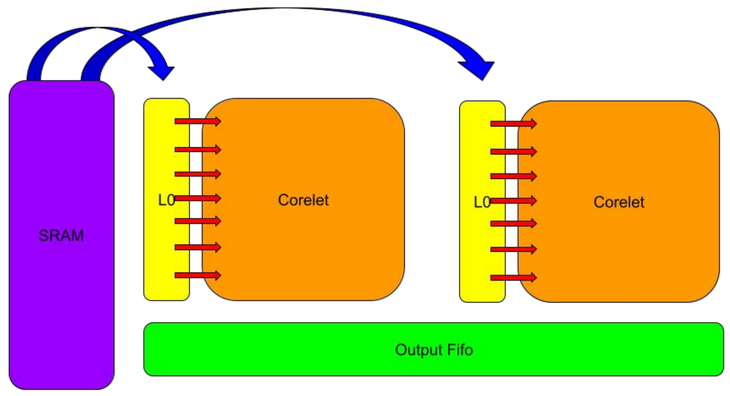
Tessier, Hugo. “Neural Network Pruning 101.” *Medium*, Towards Data Science, 13 Sept. 2021, towardsdatascience.com/neural-network-pruning-101-af816aaea61.

Alpha 2: 50% Structured & Unstructured Pruning



	Structured Pruning	Unstructured Pruning
Initial Accuracy after Pruning:	10%	37.81%
Accuracy after Fine Tuning:	91%	88%

Alpha 3: Dual-core Processor



Implemented a dual-core processor allowing for an 8 input channel and 16 output channel convolutional layer.