# Weight- and output-stationary reconfigurable 2D systolic array-based AI accelerator and mapping on Cyclone IV GX

SOC - Bei Pei, Marin Cao, Xinyuan Cai, Andy Zhang, David Wang

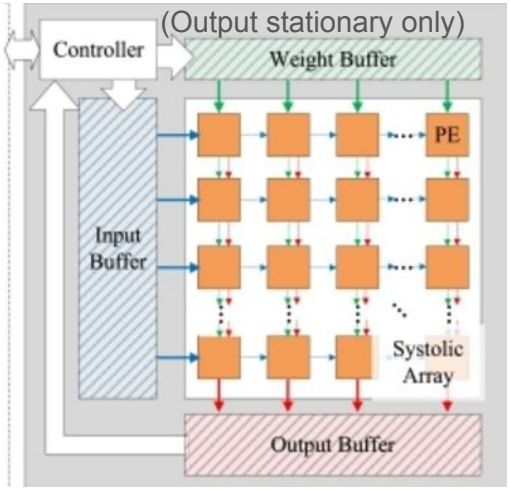
## Motivation:

As machine learning models grow in complexity, it becomes increasingly important to keep efficiency in mind. A 2D systolic is able to deliver significant performance boosts by effectively calculating MAC (Multiply and Accumulate) operations. We designed a 2D systolic array equipped with weight stationary mapping and output stationary mapping and experimented with various methods of increasing efficiency.

## Mapping on FPGA (Cyclone IV GX)

| | VGGNet |
|---|---|
| Frequency | 125.69 MHz |
| Dynamic Power | 27.4 mW |
| TOPs/W | 0.588TOPs/W |
| TOPs/s | 0.0161TOPs |
| Registers | 22,221 |
| Logical Elements | 12,098 |

## 2D Systolic Array




Reconfigurable 8 x 8 array architecture for both weight and output stationary mapping.

## Alpha 1: Resnet

Quantized and compressed convolution layer on Resnet20 to fit our 8x8 2D systolic array

## VGGNet & Resnet

4 bit activation and weight quantization for integer operations in MAC

| | Accuracy (Cifar 10) |
|---|---|
| VGGNet16 | 92.67% |
| Resnet20 | 90.94% |

## Alpha 2: Structured Pruning on VGGNet

- 40% of the output channels with the smallest L1 norm puned
- ⅜ columns in 2D systolic array can be skipped, leading to lower power usage

## Alpha 3: Tiling

64 8x8 arrays for 64x64 convolution layer in Resnet20

## Acknowledgements:

Starter code and main project outline are from Professor Mingu Kang, UCSD ECE 284.