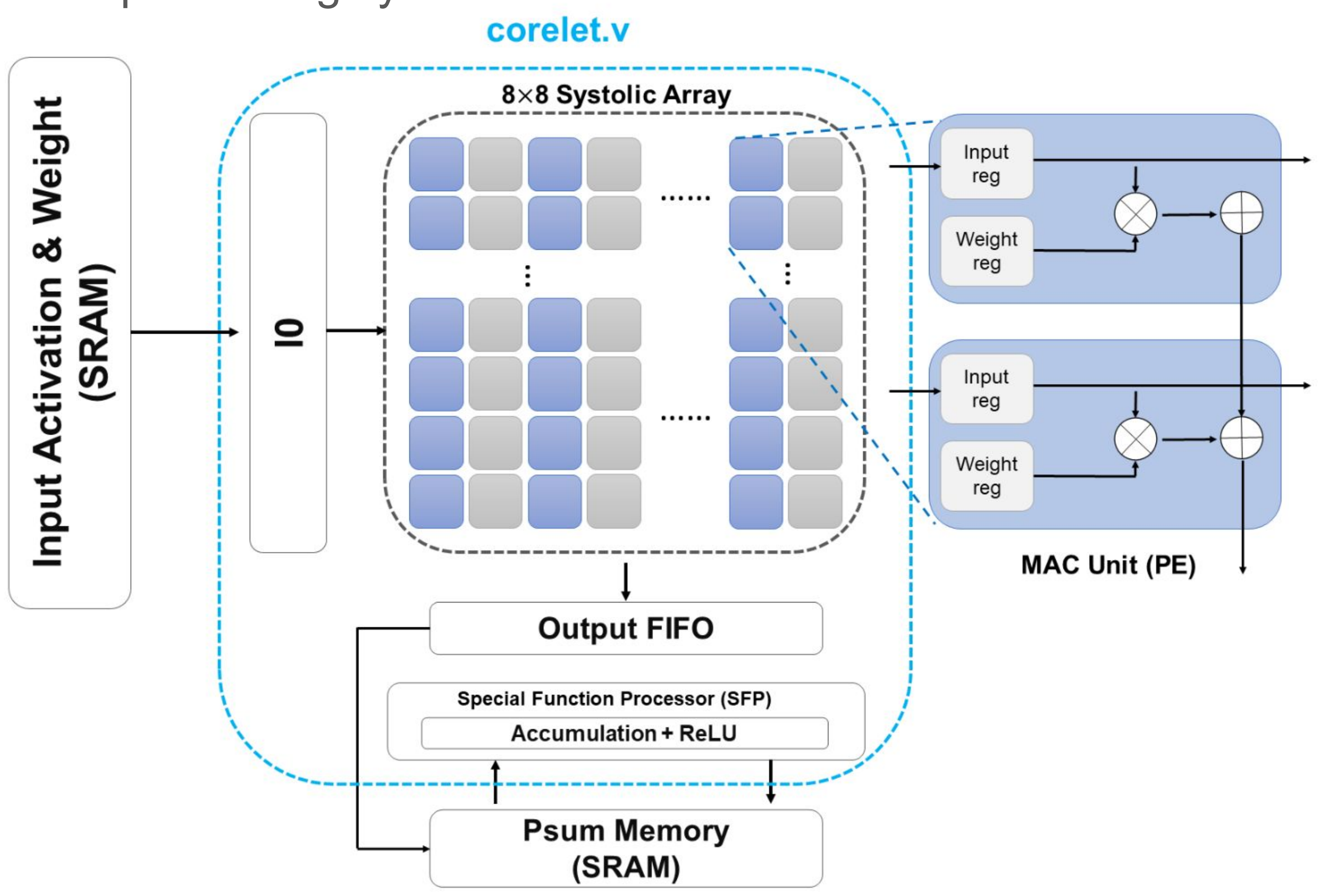
# CNN Accelerator based on Reconfigurable 2D-Systolic Array

Group Hachimi: Ziyi Xiao, Hamad Alajeel, Chengguo Yang, Yunyi Zhu, Ash Zhou

## Motivation

To push our 2D-systolic CNN accelerator toward higher performance, flexibility, and energy efficiency, we introduced three optimizations:

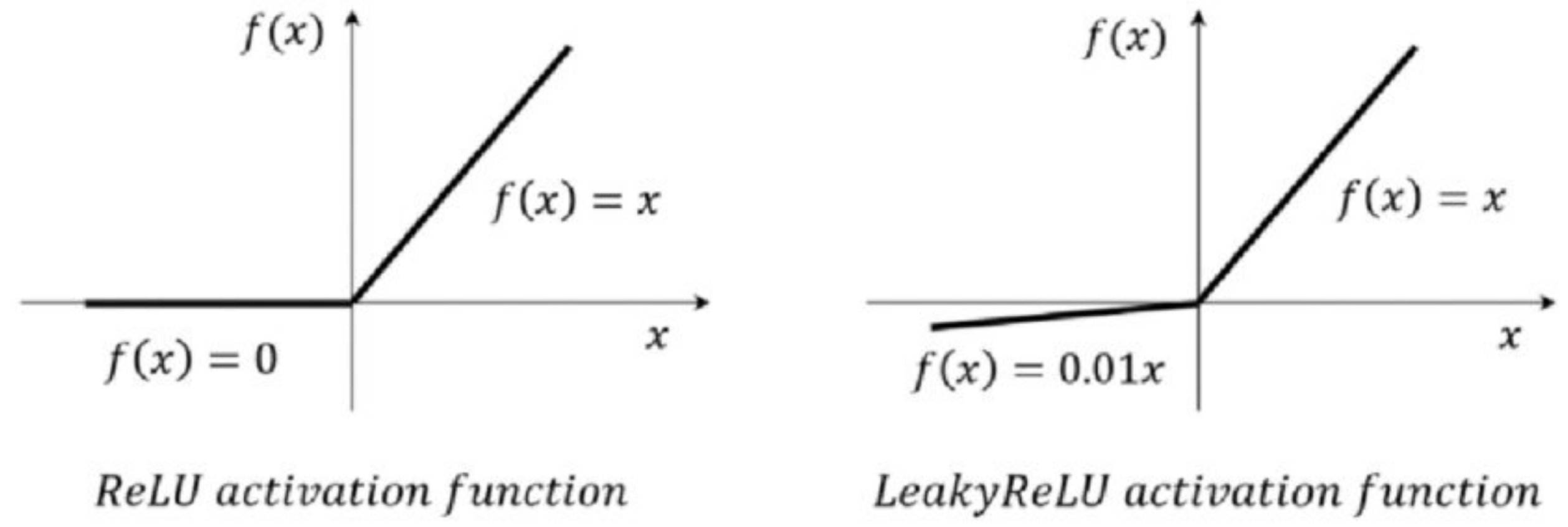
- $A_{,1}$: Versatility and reconfigurability of activation function implementation.
- $A_2$: with pruned ResNet20 for lower compute/memopy cost.
- $A_3$: with restructured execution schedule for reducing processing cycles.



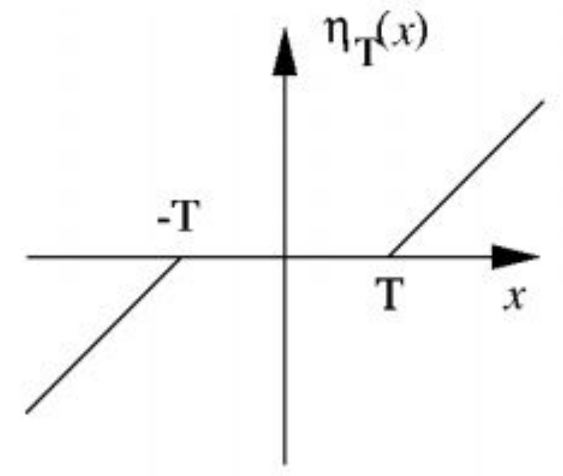
## Mapping on FPGA

| | Vanilla | Bit Reconfigurable | Weight/Ouput Reconfigurable |
|---|---|---|---|
| Total Registers7 | 12098 | 21265 | 15161 |
| Total Logical Elements | 17233 | 12451 | 23000 |
| Frequency | 127.5MHz | 90.59MHz | 88.97MHz |
| Dynamic Thermal Power | 33.48mW | 20.97mW | 27.36mW |
| GOPs | 16.32 | 11.60 | 11.39 |
| GOPs/W | 487.45 | 395 | 426.45 |

## Alpha 1: Alternative Activation Functions



To allow the NPU's we are designing to become more versatile. We have allowed the activation function implemented in the SFU's to be reconfigurable. Our SFU module is able to emulate Leaky relu and a set of parametric relu functions by using bit shifting for the negative slopes of these functions. Typically, the negative slope of leaky relu is 0.1 which can be implemented in the SFU before dequantization by bitshifting psums to the right by 3 positions. This is equivalent to division by 0.125 which is ~0.1. The error introduced by this is negligible, and the resulting output feature maps which are generated are almost equivalent to ones which had leaky relu applied after dequantization.



Soft Shrink Activation

In addition to variants of relu, our NPU is able to apply the soft shrinkage activation function which is used in denoising networks such as SCNN's and are elements of convolutional denoising networks used in medical imaging.[1] This was implemented by calculating the integer, c, such that:

$$T < \left|(c+1)\left(\frac{\alpha_{\text{act}}}{2^{\alpha_a}-1}\right)\left(\frac{\alpha_w}{2^{\alpha_w-1}-1}\right)\right| \quad \text{and} \quad T \geq \left|c\left(\frac{\alpha_{\text{act}}}{2^{\alpha_a}-1}\right)\left(\frac{\alpha_w}{2^{\alpha_w-1}-1}\right)\right| \qquad (1)$$

Implementing a multiplexer in the SFU which uses the derived value for c to apply soft shrinkage thresholding eliminates the need for post NPU activation using floating point operations and introduces negligible error while allowing for the NPU to be applied in more versatile settings, such as medical imaging.

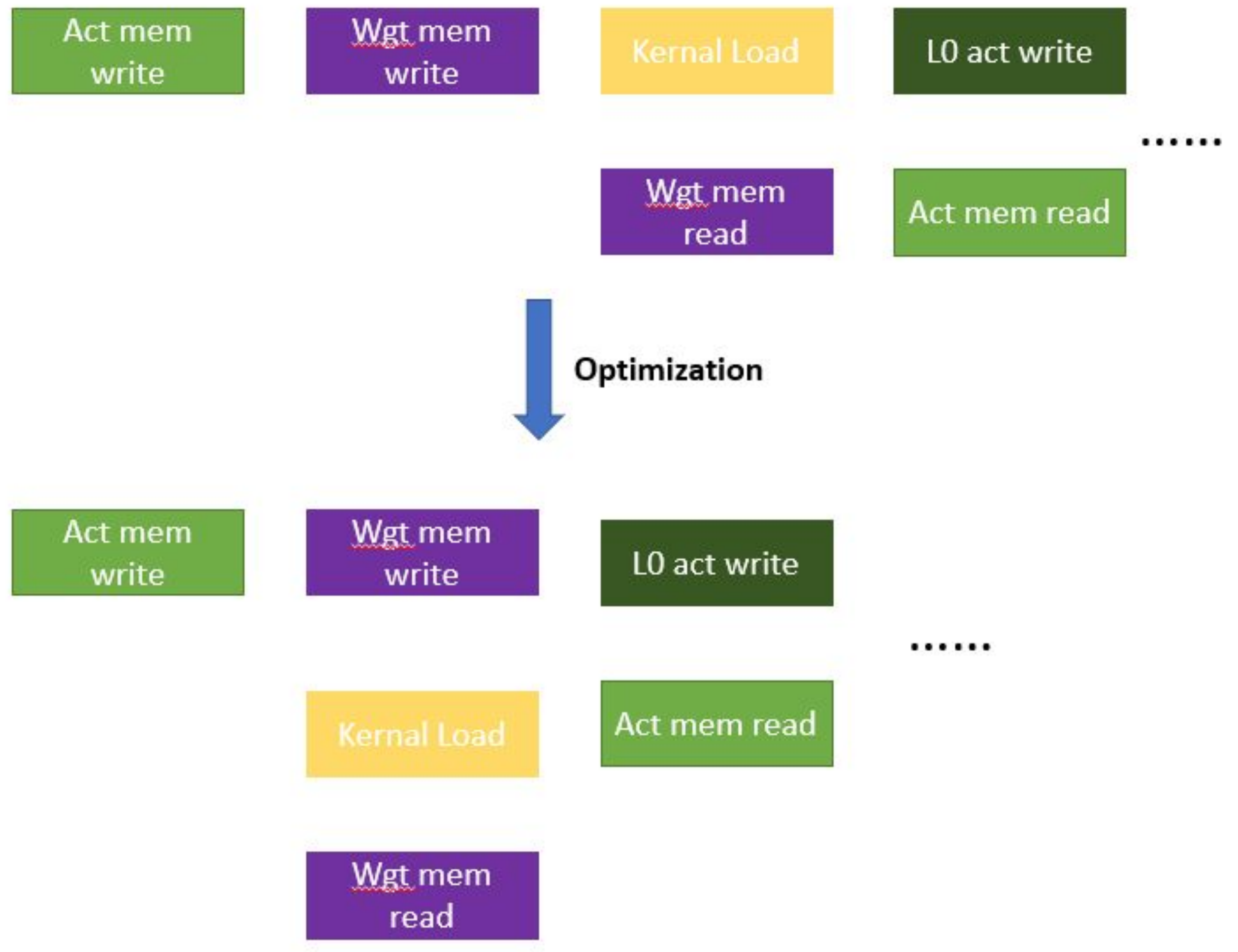
## Alpha 2: Deployment of ResNet20

4-bit QAT with 80% sparsity ResNet20 Model was trained to verify the generalization ability of 2D Systolic Array and the efficiency of pruning

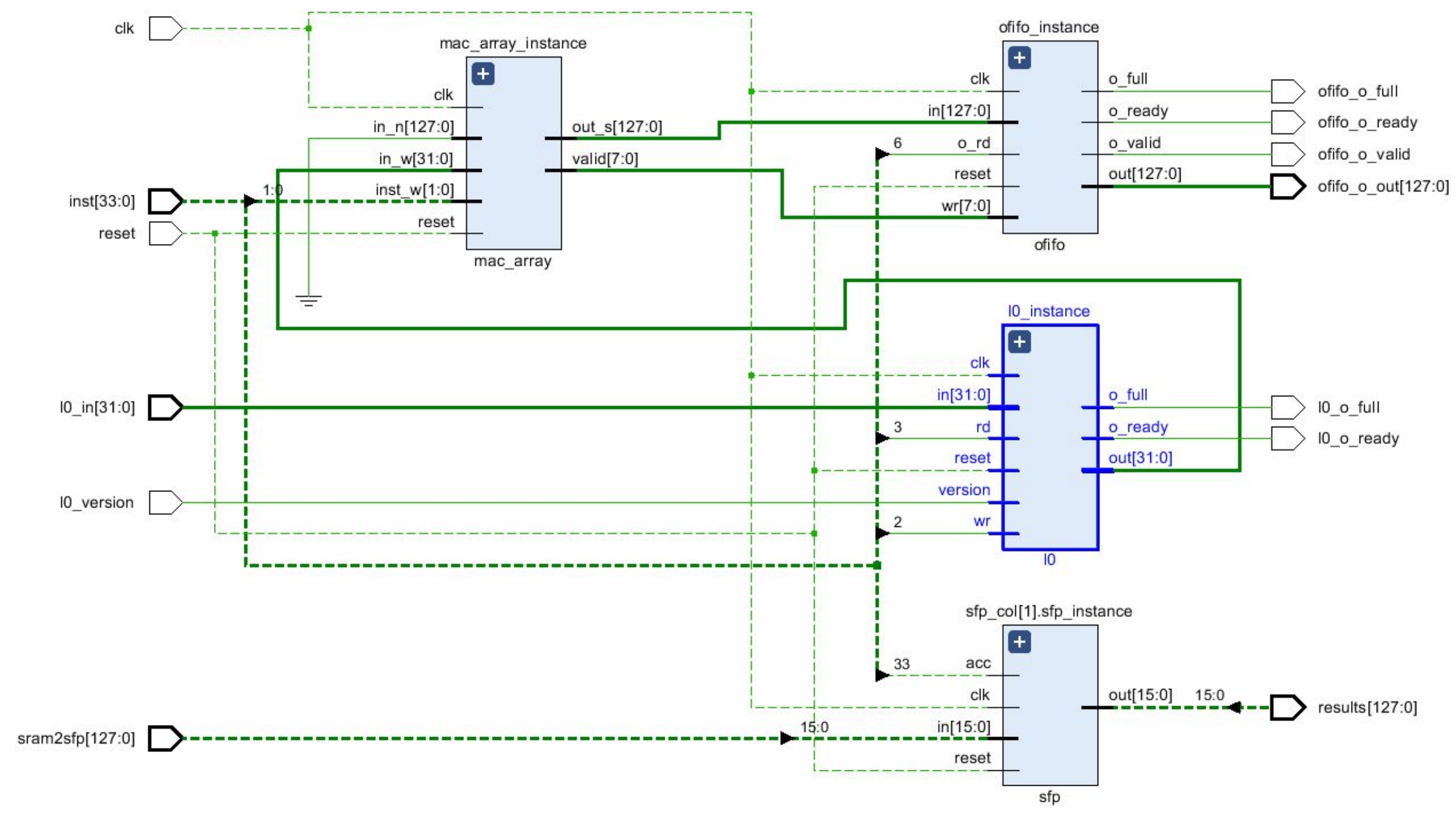
| | VGG16 | ResNet20 |
|---|---|---|
| Accracy | 90.08% | 87.59% |
| Quantization Error | 1.05e-07 | 4.31e-08 |

## Alpha 3: Processing Cycle Reduction

We optimized processing cycle by combining weight memory write to I0 and weight loading to kernel at the same time. This reduces the total duration of completing the whole operation of inference using our NPU.



## Design of corelet



## Refrences

[1] M. Kidoh, K. Shinoda, M. Kitajima, K. Isogawa, M. Nambu, H. Uetani, K. Morita, T. Nakaura, M. Tateishi, Y. Yamashita, and Y. Yamashita, "Deep learning based noise reduction for brain MR imaging: Tests on phantoms and healthy volunteers," Magnetic Resonance in Medical Sciences, vol. 19, no. 3, pp. 195–206, Aug. 2020, doi: 10.2463/mrms.mp.2019-0018.