CNNs on 2D Systolic array with optimizations

The Best Team

MingWei Yeoh, Leo Naab, Karon Luo, Ian DeGrood, Jason Phung, Allen Keng
University of California San Diego

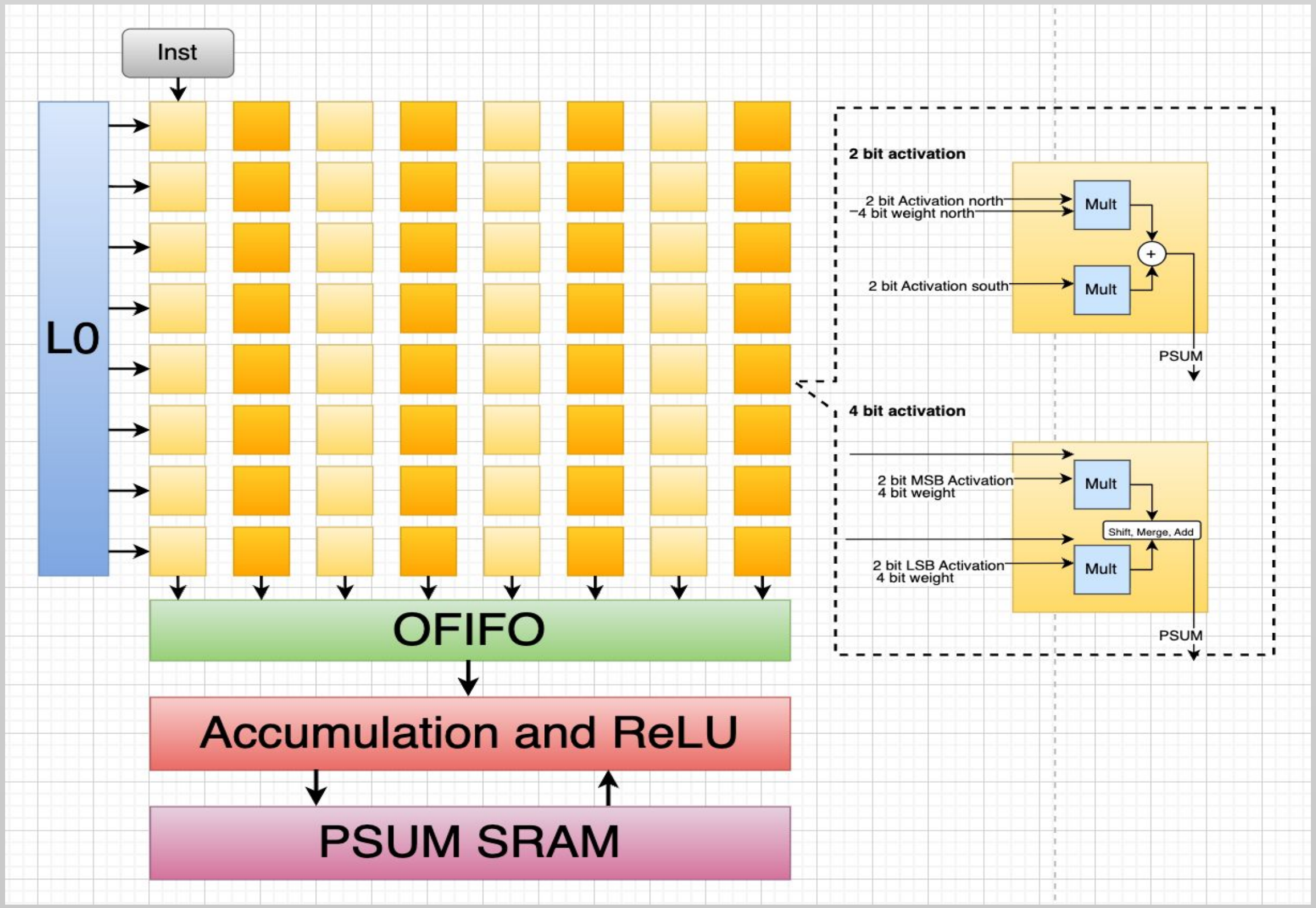
Training Results

VGG16	4-bit Activation	2-bit Activation
Accuracy	90.42 %	90.32 %
Quantization Error	3.13E-7	8.1E-6

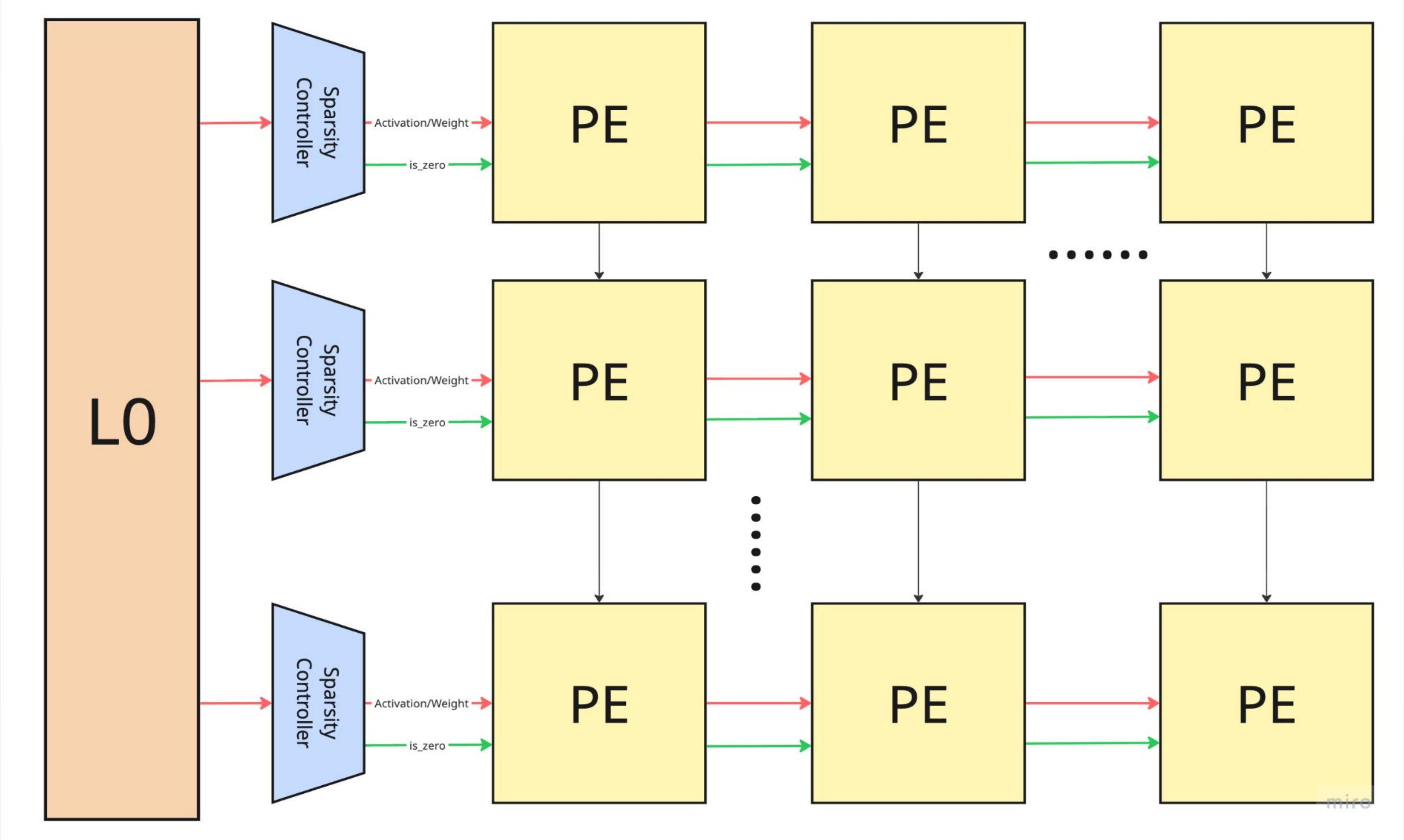
FPGA Mapping

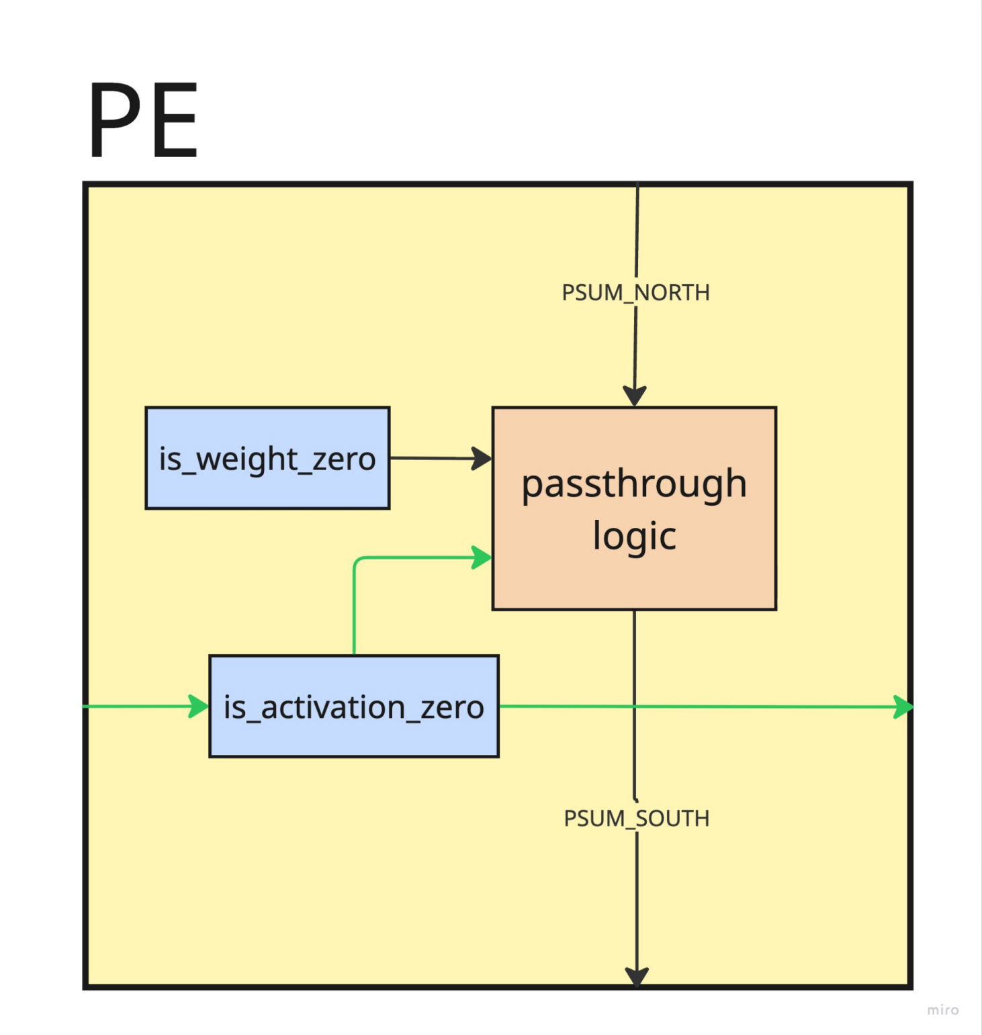
Operations	128
Frequency	129.02 MHz
Logic elements	22,845
Total: Dynamic Power	245 mW : 24 mW
GOPs	16.512
GOPs/W	67.396

2D Systolic Array Diagram



α_1- Input Sparsity Controller (Weights and Activations)





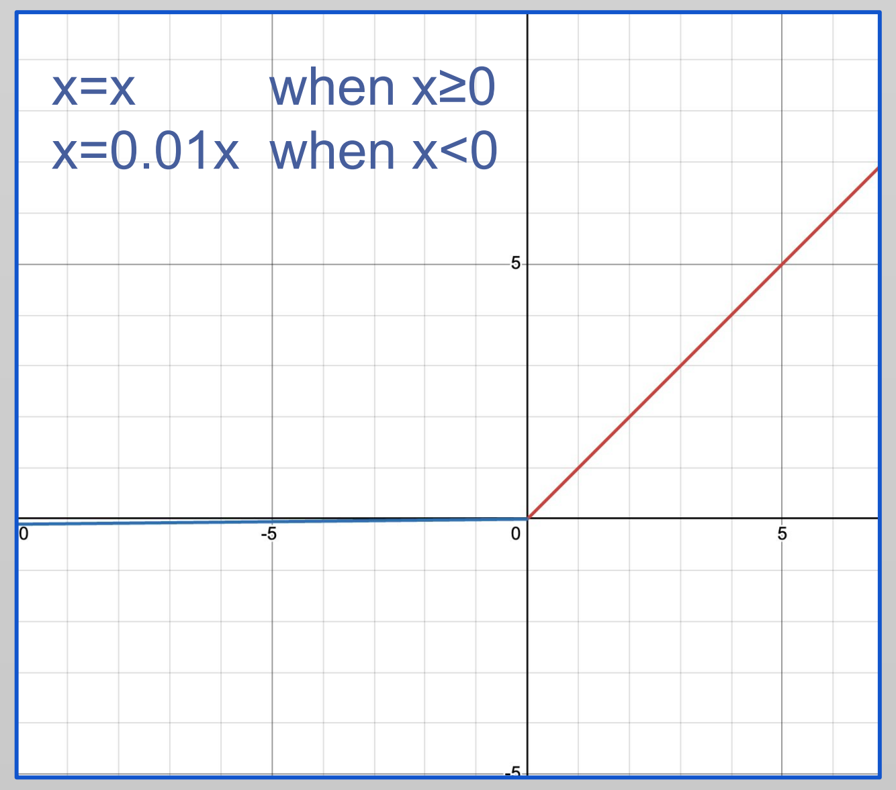
We use a sparsity controller to send a separate is_zero signal.

If the loaded weight of the current activation is zero, the PSUM gets passed from north to south without performing the power hungry multiplication and accumulation step.
We saw about a 12.5% reduction in dynamic power consumption for a given MAC tile.

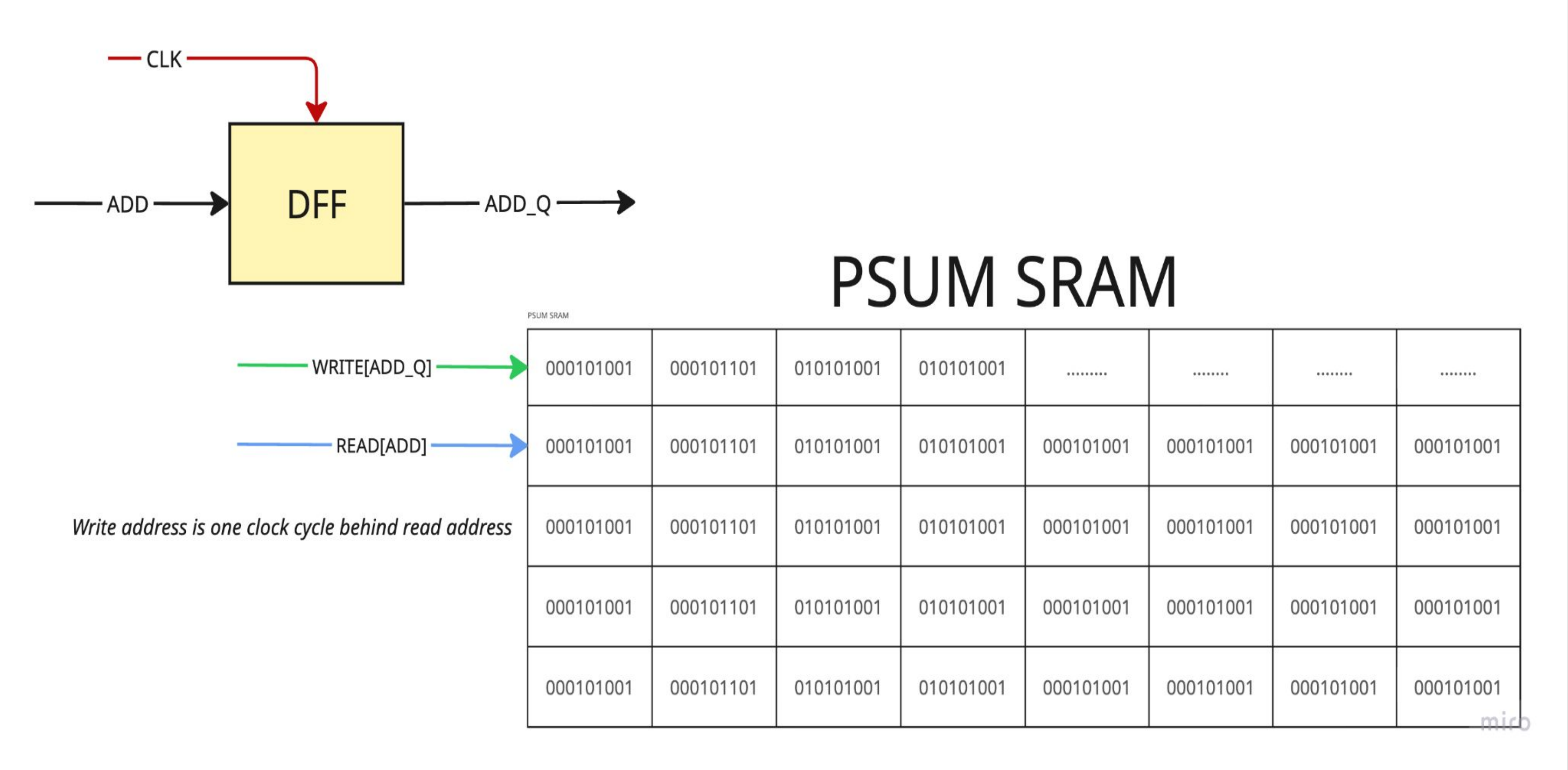
α_3- Leaky ReLU

We apply leaky ReLU activation function in the special function processor by performing a arithmetic right shift of 6 (alpha = 0.015625) to negative input instead of outputting zeros.

This improves the flow of gradients during backpropagation and leads to more stable training.

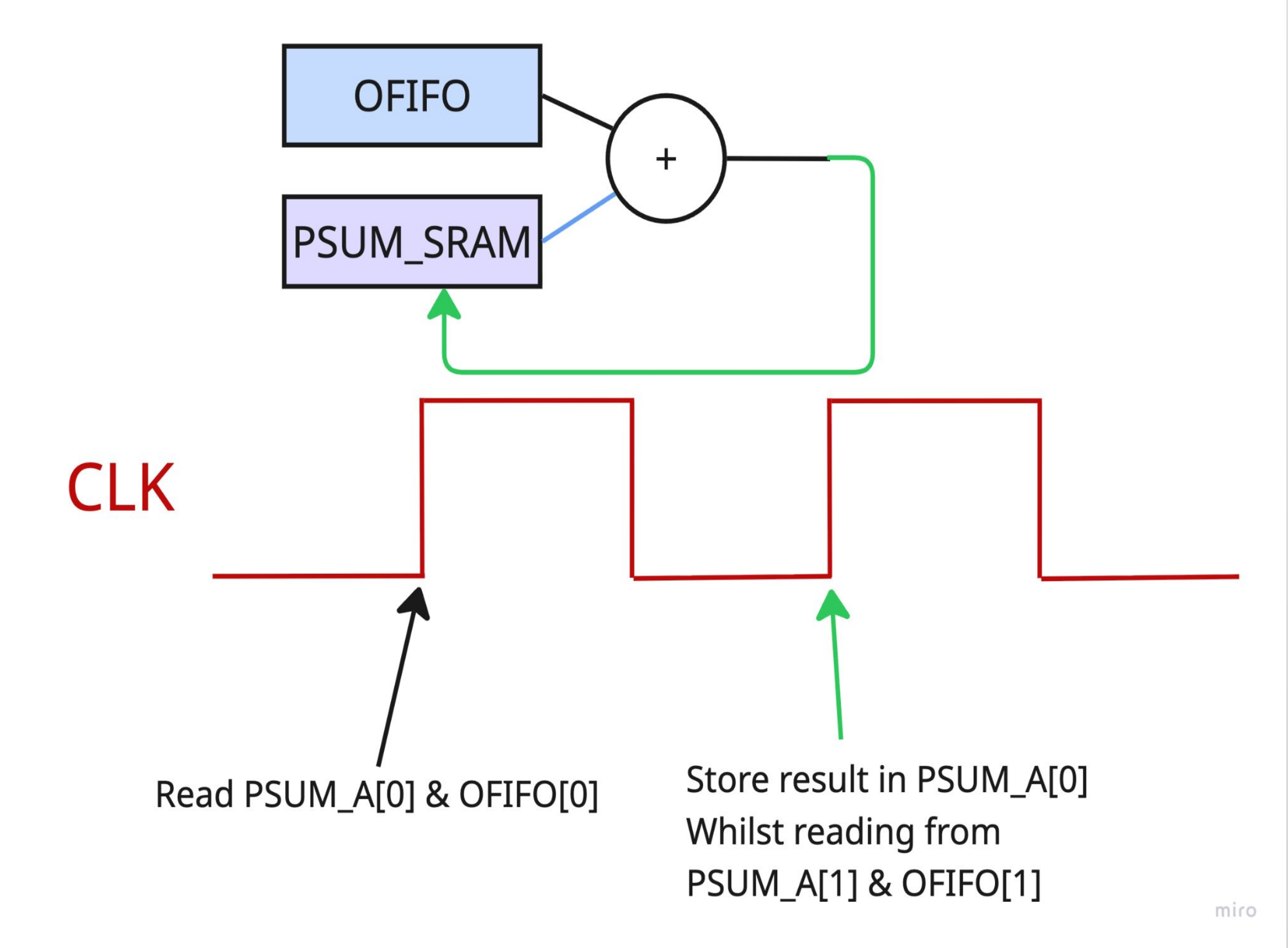


α_2- Write buffered SRAM



Write address is one clock cycle behind read address

000101001	000101101	010101001	010101001	....	....	....	....
000101001	000101101	010101001	010101001	000101001	000101001	000101001	000101001
000101001	000101101	010101001	010101001	000101001	000101001	000101001	000101001
000101001	000101101	010101001	010101001	000101001	000101001	000101001	000101001
000101001	000101101	010101001	010101001	000101001	000101001	000101001	000101001



Read PSUM_A[0] & OFIFO[0]

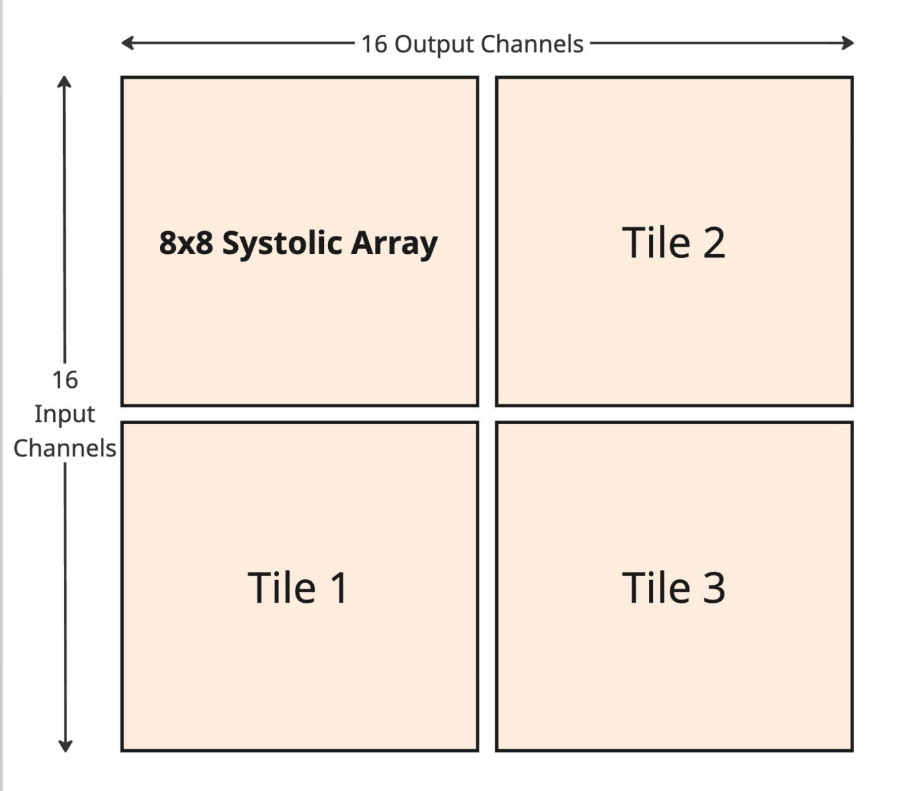
Store result in PSUM_A[0]
Whilst reading from
PSUM_A[1] & OFIFO[1]

Instead of having two SRAMs for simultaneous writing and reading, we have the write address delayed by one clock cycle to enable similar functionality using singular SRAM.

This gives you all the efficiency benefits without the complexities of having two SRAMs when time to perform the accumulation step.

α_4- Tiling

16x16 tiling allows us to perform the convolution stage on the model with less shrinking of the input and output dimensions.


16 Output Channels
16
Input
Channels
8x8 Systolic Array
Tile 2
Tile 1
Tile 3