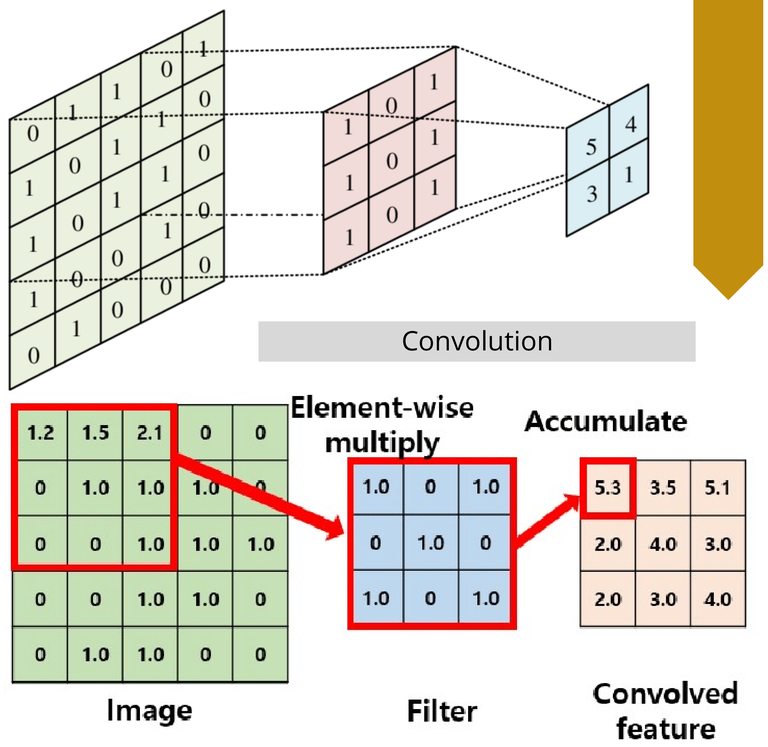
Accelerated Matrix Processor – Convolution Controller

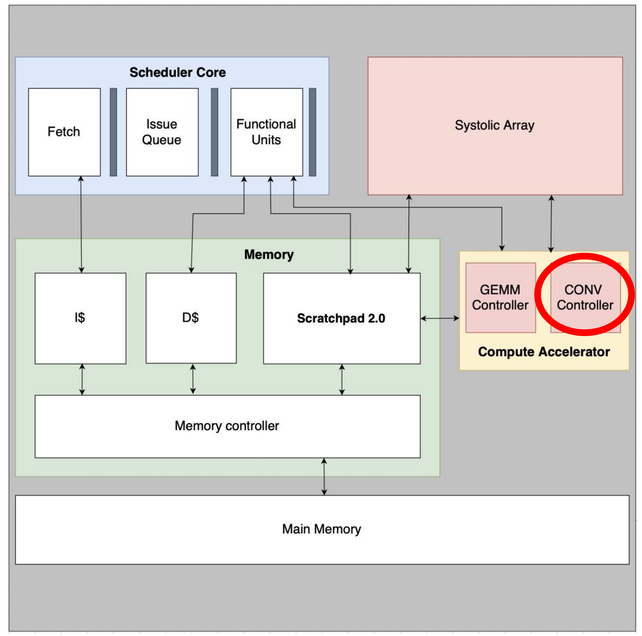
By: Saandiya KPS Mohan, Akshath Raghav Ravikiran
Mentors: Malcolm McClymont, Sooraj Chetput, Timmothy Francis Hein, Dr. Anand Raghunathan, Dr. Mark Johnson

Overview

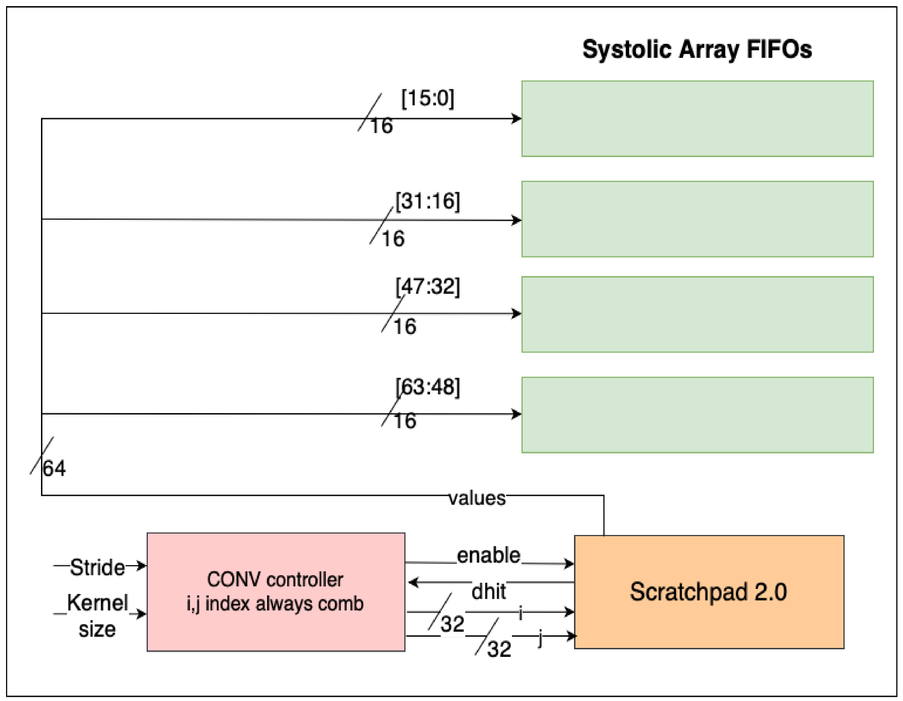
- Convolution is used in AI – CNN, GAN, encoder/decoder
- Convolution is a mathematical operation that extracts features from an input by applying a small filter (kernel) and striding across it.
- Convolution captures spatial relationships, reduces the number of parameters, and is essential for deep learning in image and signal processing.
- Existing methods – im2col in software, but has a lot of data reuse compared to GEMM
- Hence, we are designing a convolution controller and modifying AMP0 systolic array for efficient convolution operations.
- Convolution controller has to be designed to work with systolic arrays, improve efficiency for AI workloads by using spatial locality.



Convolution



Accelerated Matrix Processor - Overall Architecture



Convolution controller RTL

Method - How it Works

- Scratchpad - 2MB linearly addressed on chip memory.
- Convolution controller has an FSM to send starting address of the input matrix in the form of [row][column] to the scratchpad.
- Scratchpad sends consecutive fp16 values depending on kernel size from that starting address to systolic array.
- Convolution controller decides which FIFOs of systolic array to put each FP16 value in.
- Systolic array will perform matrix multiplication independently of the FSM.
- Some partial sum values that don't map to im2col will be ignored, while only useful partial sum values will be brought into the scratchpad.
- Reduces memory footprint compared to convolution in software because values are read from the scratchpad without duplicated accesses from DRAM.

Future Plans

- [APR] - Develop robust Python simulator with multiple channels and different kernel size to validate hardware.
- [AUG] - Implement code logic and testbench in SystemVerilog to validate the FSM.
- [OCT] - Integrate the convolution controller with other subsystems of AMP after verification.