CNN inference on 2-D Systolic Array

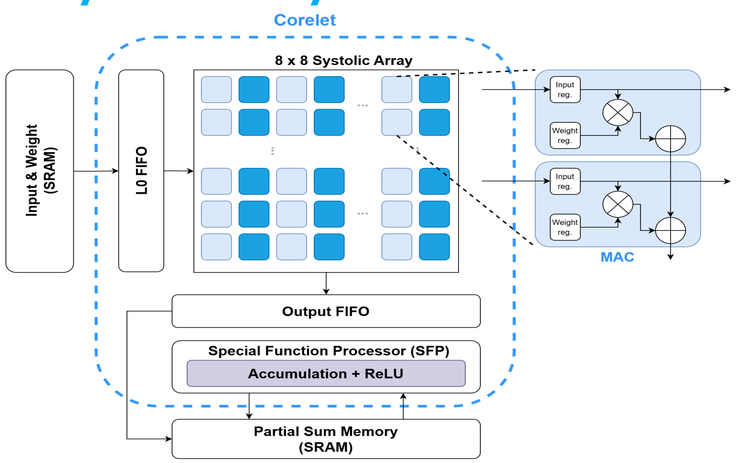
Clocked Out

Ahmet Emre Eser, Ashita Singh, Devansh Gupta, Riyansh Chaturvedi, Rohan Nafde, Shoumik Panandikar

Motivation

This project develops three specialized 2-D systolic array variants: a weight stationary design for efficient data reuse, a SIMD version supporting 2-bit and 4-bit operations, and a hybrid architecture with configurable weight and output stationary modes. Performance optimizations and comprehensive testing ensure robust acceleration across diverse neural network workloads.

2D Systolic Array



VGGNet With Quantization Aware Training

	VGGNet 16 (4 bit)	VGGNet 16 (2 bit)
Accuracy (CIFAR 10)	89.560%	85.360%

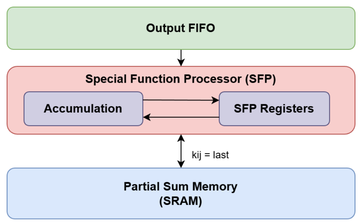
Mapping on FPGA (Cyclone IV GX)

	VGGNet 16
Total OPs	128
Frequency	130.34 MHz
TOPs	0.01668
Dynamic Power	31.24 mW
TOPs/W	0.534
Total Logic Elements	22,556
Total Registers	12,146

Base Versions

- **Vanilla (Weight Stationary):** The baseline architecture implements the fundamental Weight Stationary (WS) dataflow, primarily optimizing for weight reuse efficiency.
- **SIMD Enhanced (2/4-bit Precision):** Features a Single Instruction, Multiple Data (SIMD) data path to support 2-bit and 4-bit input precision, increasing throughput for quantized neural network operations.
- **Unified Stationary (WS & OS):** Implements a flexible structure supporting dynamic switching between Weight Stationary (WS) and Output Stationary (OS) dataflows, aiming for applicability across various computation patterns.

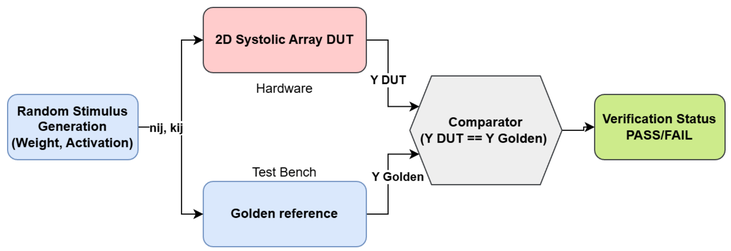
Alpha 1: Simultaneous Accumulation



We've implemented in-place accumulation through the SFP pipeline. Each SFP lane holds a dedicated accumulator register that stores the running sum of partial psums (kij) from the MAC array via OFIFO, accumulating them internally

Only the final 8×8 block psum is written to SRAM, reducing memory traffic and eliminating intermediate writes. This achieves in-place accumulation within the hardware pipeline.

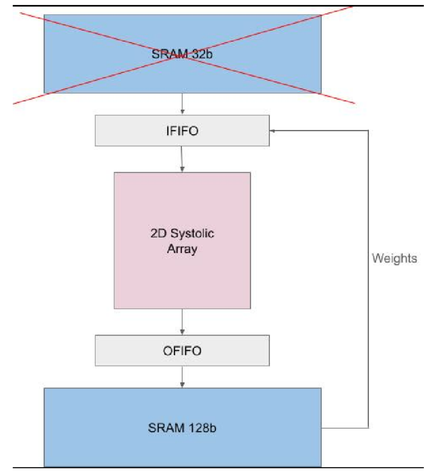
Alpha 2: Functional Verification



Concept: To ensure the integrity of our systolic array, we employ a rigorous hardware-software verification methodology.

Implementation: We utilize random stimulus testing where the array is fed random numerical inputs (weights and feature maps) spanning the full supported bit precision. A Golden Reference Model concurrently calculates the expected output. The hardware result is then compared against the Golden Result to confirm functional correctness across all operational modes.

Alpha 3: Unified Memory

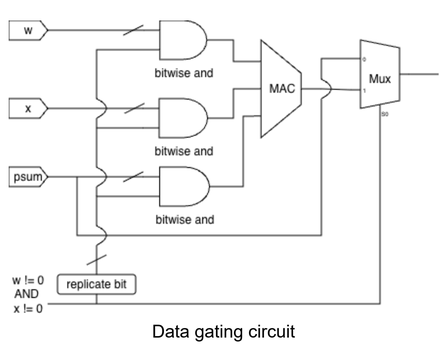


Concept: In Output Stationary mode, the Partial Sum Memory (pmem) is bandwidth-heavy during accumulation but often idle during the initial weight loading phase.

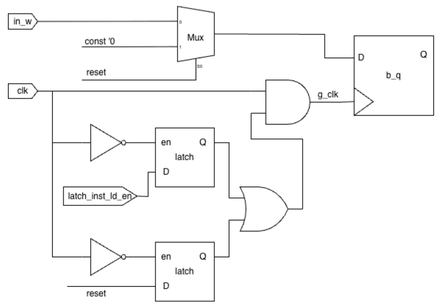
Optimization: We reuse the existing high-bandwidth pmem (128-bit) to store weights instead of adding a separate, dedicated Kernel Memory (kmem).

Benefit: This removes the need for an additional 32-bit SRAM bank entirely. Since SRAM blocks dominate the silicon area of a core, eliminating one significantly reduces the total hardware footprint and static power leakage.

Alpha 4: Circuit Level Power Optimizations



Data gating circuit



Clock gating circuit

Data Gating: When either the weight or activation is zero, the output simplifies to psum, making the MAC computation unnecessary. The MAC unit is bypassed using a MUX that forwards the psum to the output, while all inputs to the MAC are forced to zero. This prevents unnecessary switching activity within the MAC unit, saving dynamic power

Clock gating: Weight and load enable registers hold their values stable for long periods. When a register's load enable is inactive, clock gating prevents the flip-flops from toggling, avoiding the charging and discharging of internal capacitances, hence saving dynamic power.