RTL Implementation and Verification of an FP4 Systolic Array for Edge AI Workloads

Daniel Pace-Farr, Computer Systems Engineering
Mentor: Dr. Deliang Fan, Associate Professor
School of Electrical, Computer and Energy Engineering

Research Question

Can a novel 4-bit floating-point systolic array design be functionally verified to produce correct matrix multiplication results across both E3M0 and E2M1 formats?

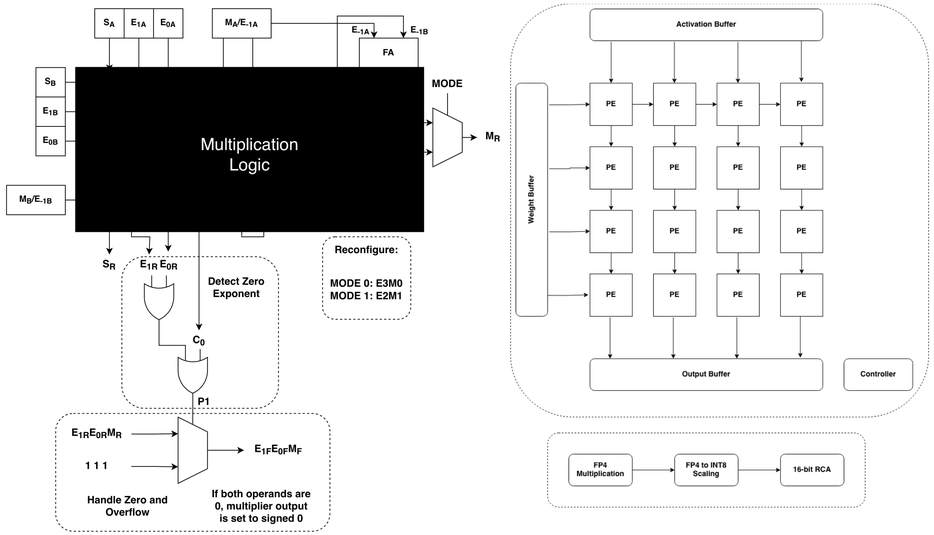
Introduction

Modern AI models rely on millions of multiply-accumulate (MAC) operations for training and inference. Quantization reduces operand bit widths to lower-precision formats like 4-bit floating-point (FP4) which drastically reduces the area and power required per operation. This study functionally verifies a novel FP4 MAC design for hardware acceleration.

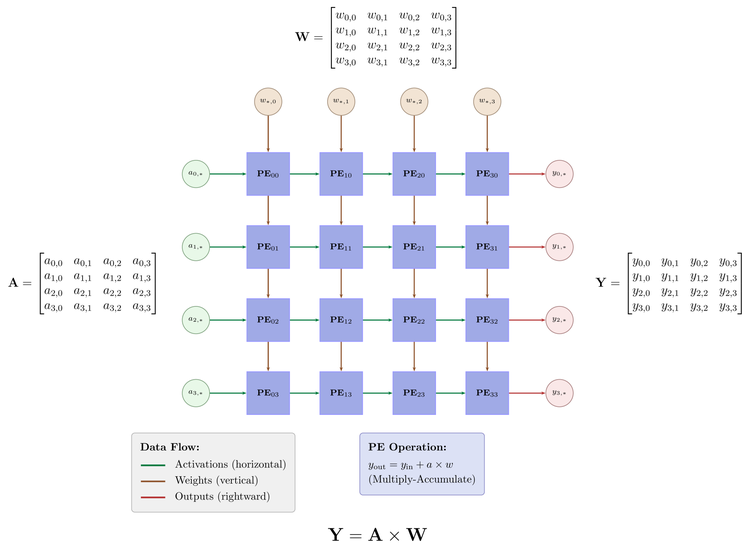
Methodology

This project verifies an FP4 systolic array accelerator design for efficient MAC operations. The architecture supports two FP4 formats: E3M0 (1 sign + 3 exponent bits) and E2M1 (1 sign + 2 exponent + 1 mantissa bit) for precision tradeoffs between dynamic range and granularity. The core design is a 4×4 output-stationary systolic array, where each processing element (PE) performs FP4 multiplication, converts the result to INT8 via lookup table, and accumulates partial products in a 16-bit integer register. All 16 output accumulators are valid 7 cycles after the first data enters the array.

Verification was performed using exhaustive and functional testbenches in SystemVerilog and simulated with ModelSim. The design was tested against all 256 input combinations for each FP4 format mode.



Top Left: *Reconfigurable FP4 multiplier supporting E3M0 and E2M1 modes via mode-selected multiplexers with zero-detection and overflow saturation handling.*
Top Right: *4x4 output-stationary systolic array with activation and weight buffers feeding into processing elements that each perform FP4 multiplication, FP4-to-INT8 scaling, and 16-bit accumulation.*
Bottom: *Data flow of the 4x4 output-stationary systolic array computing Y = AW, with activations propagating horizontally and weights propagating vertically through MAC processing elements.*



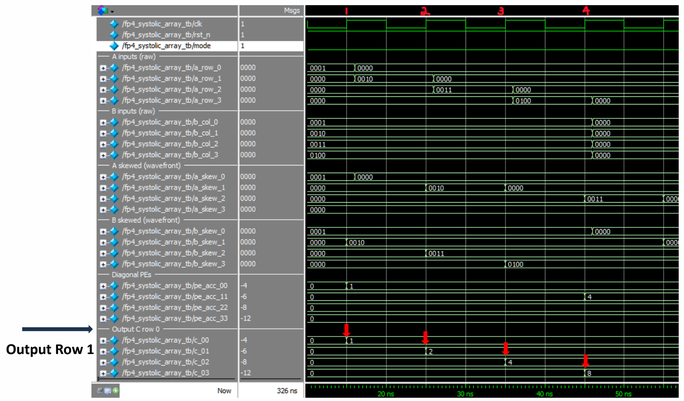
Results

All hardware modules were verified through exhaustive and functional testbenches totaling 678 individual assertions, with a 100% pass rate across both FP4 formats. The FP4 multiplier was tested against all 256 possible input combinations per mode, confirming correct sign, exponent, and mantissa computation as well as proper zero-detection and overflow saturation behavior.

Functional Verification Results

Module	Tests	Pass Rate
FP4 Multiplier (E3M0 + E2M1)	512	100%
Processing Element	19	100%
4 × 4 Systolic Array	48	100%
INT8 Conversion	51	100%
End-to-End	48	100%
Total	**678**	**100%**

The systolic array waveforms confirm correct diagonal wavefront propagation, with the figure below showing the sampled accumulator outputs of row 0 settling one cycle apart.



Waveform of accumulator outputs showing the staggered wavefront produced by the 4x4 systolic array.

FURI

Ira A. Fulton Schools of Engineering
Arizona State University