Design Documentation: INT8 Systolic Array AI Accelerator

Rajdeep Saha
IIIT Bangalore (ECE)

December 2025

1 Project Overview

This document details the design and implementation of a domain-specific hardware accelerator for Neural Network inference. The core architecture is a 4x4 Systolic Array utilizing a Weight-Stationary dataflow. The design targets high-efficiency INT8 (8-bit integer) matrix multiplication on Xilinx FPGAs.

The system was implemented in Verilog HDL, simulated using Vivado 2022.2, and synthesized targeting the Kintex-7 (XC7K70T) FPGA.

2 Hardware Architecture (RTL Description)

The design consists of four primary Verilog source files, organized hierarchically.

2.1 Processing Element (`pe.v`)

The fundamental computational unit of the array.

- **Functionality:** Performs the Multiply-Accumulate (MAC) operation:

$$P_{sum} = P_{in} + (Activation \times Weight)$$

- **Data Types:** Inputs and Weights are 8-bit signed integers (INT8). Accumulation is performed in 24-bit to prevent overflow during deep learning operations.
- **Logic:** Includes internal registers to store stationary weights. It supports a `load_weight` control signal to switch between "Weight Loading Mode" and "Compute Mode."

2.2 Data Skew Buffer (`skew_buffer.v`)

To ensure correct systolic synchronization, input data must arrive at different columns at different times.

- **Functionality:** Implements a variable delay line for the input vector.
- **Delays:**

  - Column 0: 0 cycles delay.
  - Column 1: 1 cycle delay.
  - Column 2: 2 cycles delay.
  - Column 3: 3 cycles delay.

- **Note:** Includes a synchronous `rst` signal to clear pipeline registers to zero, preventing "X" (unknown state) propagation during simulation.

2.3 Systolic Array Core (`systolic_array.v`)

The interconnect module that instantiates the 4x4 grid of PEs.

- **Topology:** 2D Mesh.
- **Connections:**
  - **Vertical:** Activation data flows North → South.
  - **Horizontal:** Partial sums flow West → East.
- **Connectivity:** Uses `generate` loops to automatically wire the 16 PEs, mapping top-level ports to the mesh boundary.

2.4 Top-Level Accelerator (`matrix_accelerator.v`)

The wrapper module that integrates the Skew Buffer and the Systolic Array.

- **Critical Logic:** Implements a **Bypass MUX** for weight loading.
- **Mechanism:** When `load_weight` is HIGH, the input data bypasses the Skew Buffer. This ensures that weights reach all columns simultaneously (synchronous load), preventing data scrambling. When LOW, inputs pass through the Skew Buffer for correct diagonal wavefront computation.

3 Verification Environment

Verification was performed using two testbenches to validate functionality at both the unit and system levels.

3.1 Unit Testing (`tb_pe.v`)

- **Objective:** Verify the signed arithmetic and register logic of a single PE.
- **Test Cases:** Checked positive multiplication, accumulation over multiple cycles, and handling of negative weights (2's complement).
- **Status:** PASSED.

3.2 System Verification (`tb_matrix_accelerator.v`)

- **Objective:** Verify the full matrix multiplication logic.
- **Methodology:**
  1. Loaded a 4×4 Identity Matrix into the array.
  2. Streamed an input vector $A = [1, 1, 1, 1]$.
  3. Observed the Output Vector C.
- **Expected Result:** $C = A \times I = [1, 1, 1, 1]$.
- **Status:** PASSED (Waveform confirms outputs $01, 01, 01, 01$).

4 Physical Constraints

4.1 Timing Constraints (`timing.xdc`)

A timing constraint file was created to define the target operating frequency for Synthesis.

- **Target Clock:** 100 MHz (10.00 ns period).

5 Results & Analysis

5.1 Functional Output

The behavioral simulation confirms bit-accurate functionality. The output signal `result_out[95:0]` correctly displays four distinct 24-bit segments, each holding the value `1` (Hex: `01`), validating the Identity Matrix multiplication test.

5.2 Synthesis Results (Kintex-7)

The design was synthesized on the XC7K70T FPGA.

Metric	Value	Analysis
Target Frequency	100 MHz	Base Constraint
Worst Negative Slack (WNS)	+5.137 ns	Timing Met with large margin
Max Frequency	**205 MHz**	Derived from Slack
Slice LUTs	1487	Low utilization ($< 3\%$)
Slice Registers	624	Efficient Pipelining
Total Power	0.123 W	Suitable for Edge AI

Table 1: Post-Synthesis Performance Metrics

6 Conclusion

The project successfully demonstrated a functional INT8 Systolic Array Accelerator. The design meets the 100 MHz timing requirement with significant headroom (capable of 205 MHz) and maintains a low area footprint, making it suitable for deployment in power-constrained embedded systems.