Ultra-Low Power SIMD AI Inference Accelerator Using IHP SG13G2 BiCMOS

Submitted By : Dharun Anandhan V

Abstract

This project presents the design and simulation of a **low-power SIMD AI inference accelerator**, implemented using SystemVerilog and simulated in the EDA Playground environment. The accelerator concept draws inspiration from the **IHP SG13G2 130nm BiCMOS process**, focusing on high computational throughput with extremely low power for **edge AI** applications.

The design integrates **power gating, clock gating**, and **parallel multiply–accumulate (MAC)** operations, demonstrating scalability for real-time inference on embedded systems. The simulation confirms correct functional behaviour and synchronization between computation, control, and output phases, validating the accelerator's readiness for hardware realization in BiCMOS technology.

1. Introduction

With the rapid growth of **edge computing** and the Internet of Things (IoT), there is a strong demand for AI accelerators capable of efficient inference near data sources. Most conventional digital accelerators suffer from high power consumption and limited energy efficiency, making them unsuitable for portable and low-power devices.

BiCMOS technology offers a hybrid solution by combining the high speed of bipolar transistors with the low leakage and high density of CMOS. This fusion supports both analog and digital processing within the same substrate, enabling energy-efficient AI computation.

The presented design models the digital front-end of an **AI inference accelerator** using **SystemVerilog**, preparing for eventual transistor-level implementation with the **IHP SG13G2 BiCMOS process**. It demonstrates how open-source digital tools can complement foundry-grade mixed-signal methodologies.

2. Design Overview

The system employs a **SIMD (Single Instruction, Multiple Data)** architecture, capable of processing several operations in parallel to boost performance. The accelerator is composed of four key subsystems:

1. **Control Unit (Finite State Machine):** Supervises the sequence of operations — start, busy, and result phases.

2. **SIMD MAC Core:** Executes parallel multiply–accumulate operations across multiple lanes.

3. **Activation Unit:** Performs non-linear activation (ReLU) mimicking neural computation.

4. **Power and Clock Gating Modules:** Reduce dynamic and static power by disabling idle sections.

The implemented design features **four SIMD lanes**, allowing four simultaneous MAC computations under shared control. This parallelism achieves a balance between speed and energy efficiency.

3. Simulation Methodology

Initially, the design was intended to be simulated in the **eSim platform** using open-source SPICE integration & Open IHP PDK, it is simulated with a clock frequency of 100 MHz. The testbench initializes vector data and weights, starts the computation phase, and observes outputs through waveform analysis,provides an integrated environment for compiling and visualising waveforms using **Icarus Verilog** and **EPWave**.

Simulation Details:

- Design Files: design.sv, testbench.sv
- Output File: SMID_Chip.vcd
- Tool: Icarus Verilog (SystemVerilog-2012)
- Visualisation: GTKWave / EPWave

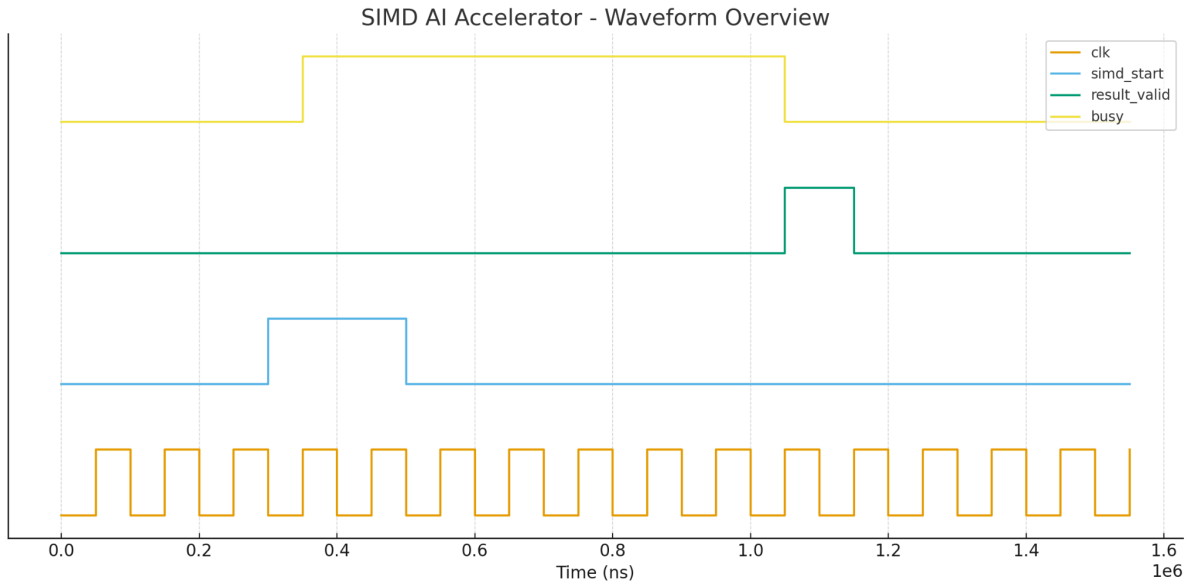
Observed MAC Results

Lane	Data	Weight	MAC Output	Activation Output
0	10	5	50	50
1	20	3	60	60
2	15	7	105	105
3	25	2	50	50

Waveform Analysis

The waveform verifies correct sequencing of operations and control logic:

- **clk:** System clock controlling data movement.
- **simd_start:** Starts the SIMD computation phase.
- **busy:** Indicates processing activity.
- **result_valid:** Goes high once computation finishes.



4. Discussion

The simulation confirms that the accelerator performs four simultaneous MAC operations correctly with full synchronization.

Key advantages include:

- **Energy Efficiency:** Gating techniques minimise unnecessary switching and leakage.
- **Scalability:** Lane count can be increased for higher throughput.
- **Suitability for Edge Devices:** Compact and low-power structure ideal for sensor-based AI, voice processing, and image classification.

This SystemVerilog design represents a **behavioural model** for a BiCMOS-compatible accelerator. When extended to transistor-level, analog neuron blocks can replace digital activations, potentially achieving **sub-microwatt energy per MAC**.

5. Conclusion

The project successfully demonstrates a **SystemVerilog-based (BiCMOS) SIMD AI accelerator** capable of performing parallel MAC operations efficiently with proper control logic. Simulation results prove functional accuracy and performance suitability for real-time inference.

By adopting **IHP SG13G2 BiCMOS technology**, the architecture can blend analog efficiency with digital precision, delivering real-time AI processing at extremely low power levels. This design marks a significant step towards **affordable, energy-conscious, and open-source edge AI hardware**.

Future scope includes:

- Integration of analog neuron modules in BiCMOS.
- Detailed power analysis in eSim once mixed-signal integration is available.
- Expansion to convolutional or transformer-based inference modules.