Special Problem (ECE 8903) Final Report

Architectural design of a MAC unit for a Processing Element

Kaul, Ankit
GTID: 903143500

Recent advances in machine learning with neural networks has led to breakthroughs in image processing, speech recognition, game development, and healthcare. With the recent shift in hardware architecture towards heterogenous multi-cores, having a mix of cores and accelerators, improving the design of existing hardware accelerators is of prime importance to increase the application scope of these devices.

One of the important challenges in designing the architecture of hardware accelerators is to ensure the best possible tradeoff between operational flexibility and efficiency. An architectural design for a deep learning accelerator is being developed at the GREEN lab at Georgia Tech. which is aimed to be more energy efficient than other current designs. Accelerators have three important architectural blocks and peripherals: Processing Elements (PE), off-chip DRAM (Dynamic Random Access Memory), and an interface between the PE and DRAM. One of the possible improvement opportunities is to design an efficient Processing Element (PE). The PE primarily consists of a Memory block (a flip flop or a Block RAM) and a MAC (Multiplier–Accumulator).

Approach: This project encompassed the design a behavioral model of a MAC unit. The main aim is to optimize the MAC unit for increased speed and efficiency. The design of a MAC was realized by using a add and shift based design. The unit utilizes a shift register, a 2:1 multiplexer, a 32-bit ripple carry adder, and a 64-bit shift register based accumulator (product) register. The behavioral model was designed using Verilog and simulations were performed in ModelSim. The current design works for 32-bit fixed point operations, with the product stored into a 64-bit register.

Currently, the MAC design needs to be implemented on Vivado to check for area, and power consumption. This design will further be integrated with other components of the accelerator, after modifications to improve the speed to power ratio. Most commercial MAC designs are clock driven.

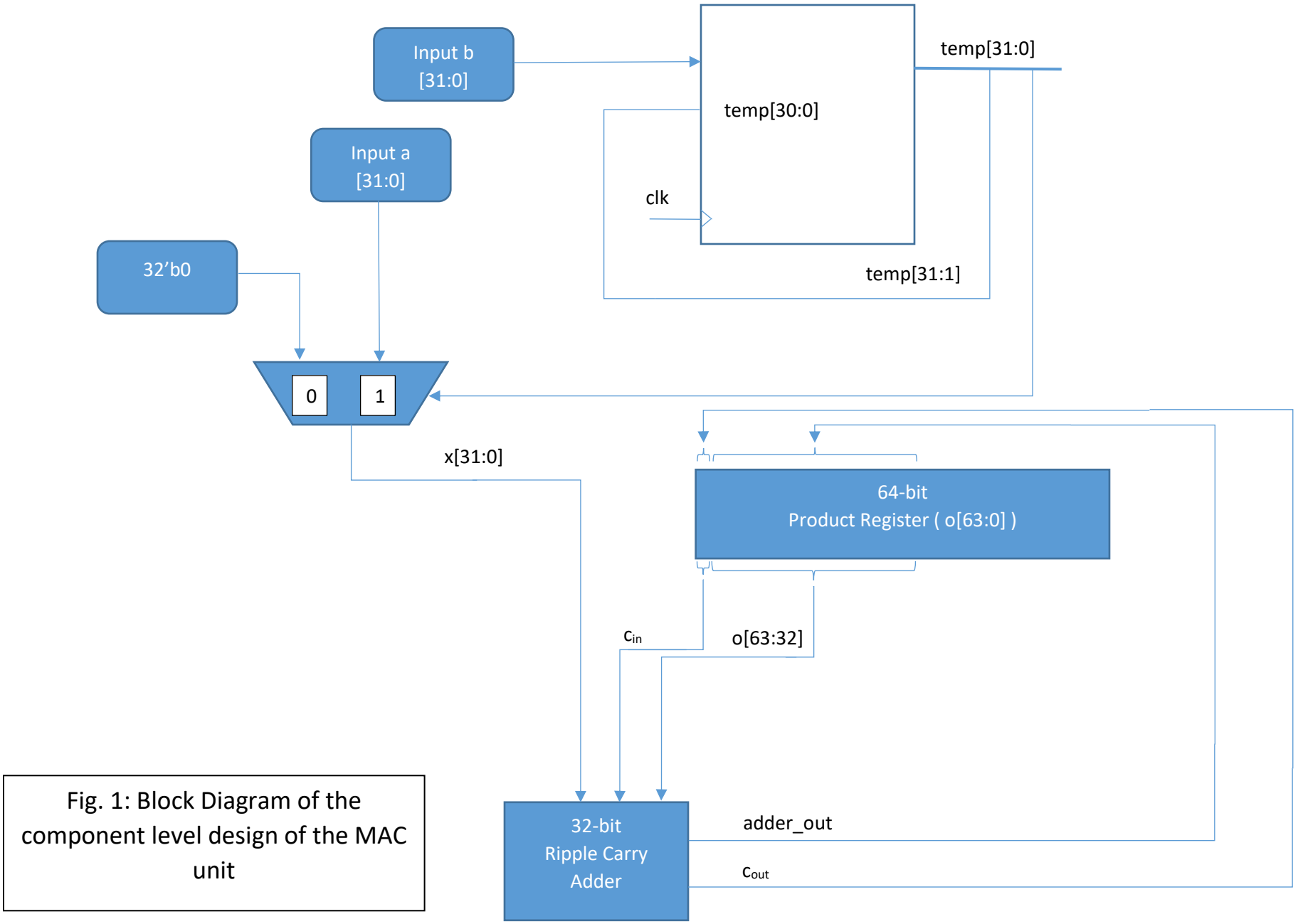

Fig. 1: Block Diagram of the component level design of the MAC unit

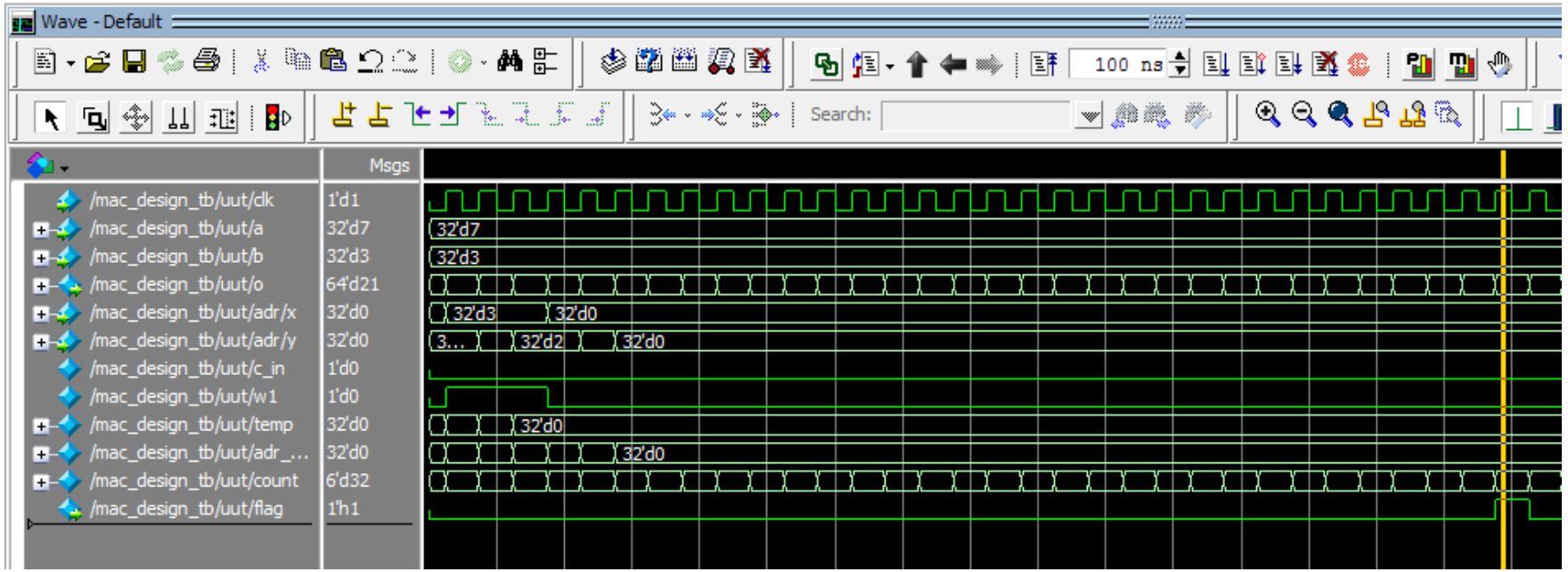

Fig. 2: Simulation output of the MAC unit in ModelSim showing the addition of two 32-bit fixed point numbers

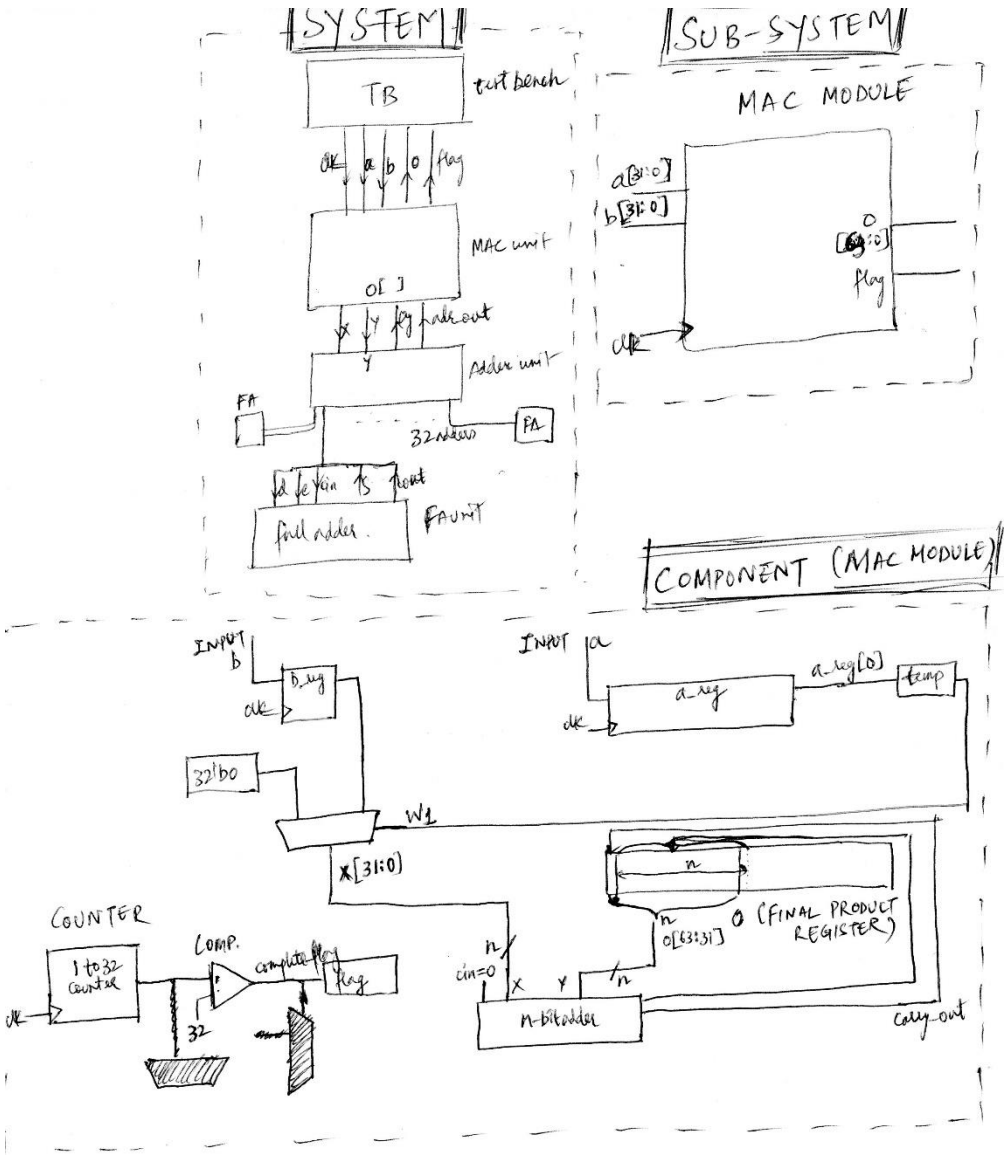

Fig. 3: System, Sub-system and Component level block diagram of the MAC unit

Further work: This design will next be modified for 32-bit floating-point operations. Subsequent efforts will focus on improving the throughput of the system by:

1. Implementing an array based multiplier, so that the product is computed in just one clock cycle.
2. Improving the interface between the PE and the off-chip DRAM.

Deliverables: Power, Area, and Timing measurements from Vivado, Commented RTL code.