MAC Accelerator Project

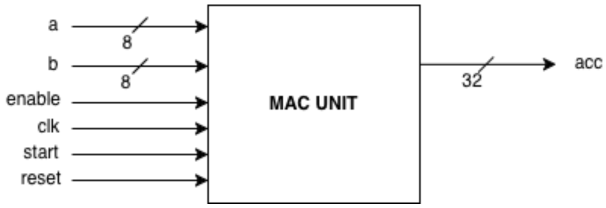
Evan Wong

Introduction:

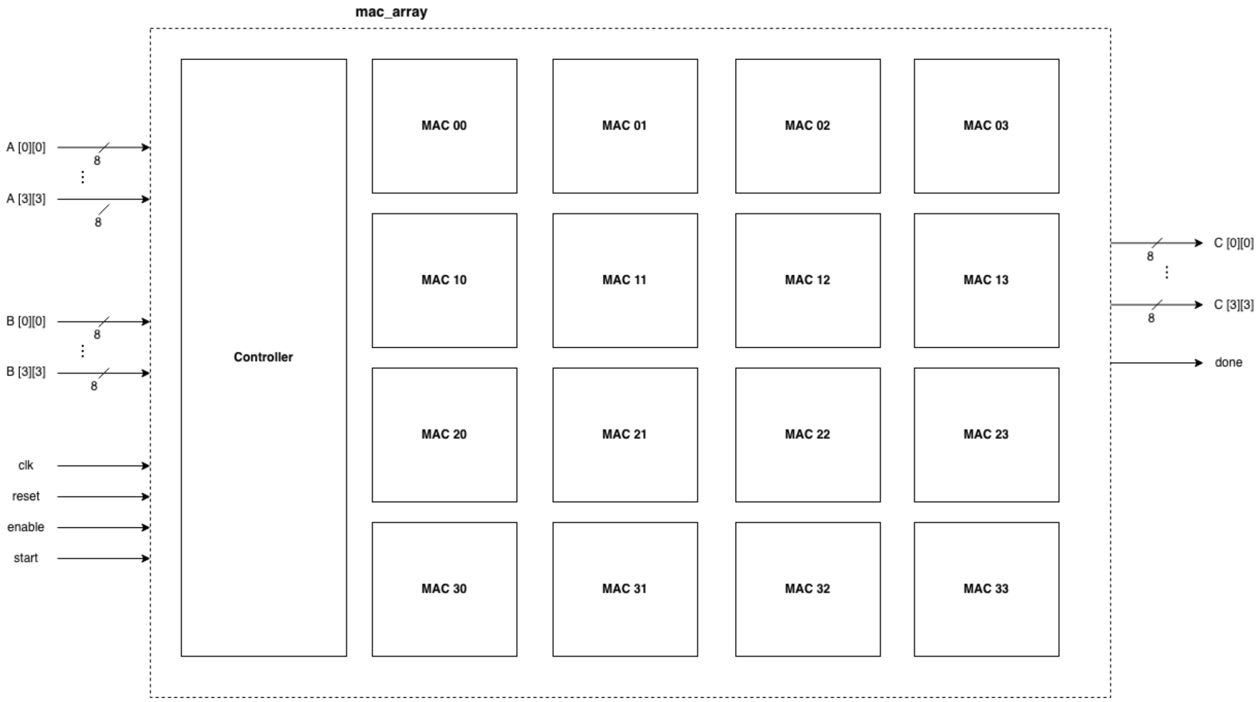
In the spirit of the “AI Revolution,” I started this project to explore the capabilities of parallel processing systems, such as GPUs, by comparing the performance of custom GPU accelerators to traditional single-pipeline processors. The goal was to design a MAC accelerator in Verilog HDL that would perform matrix multiplication. I began by constructing individual MAC units that would compute a respective cell in the resulting matrix. The aggregate system had $n \times n$ MAC units and a controller that would feed values. Each MAC unit corresponded to a specific cell in the output matrix. Computing the four MAC required four clock cycles.

Design Specifications:

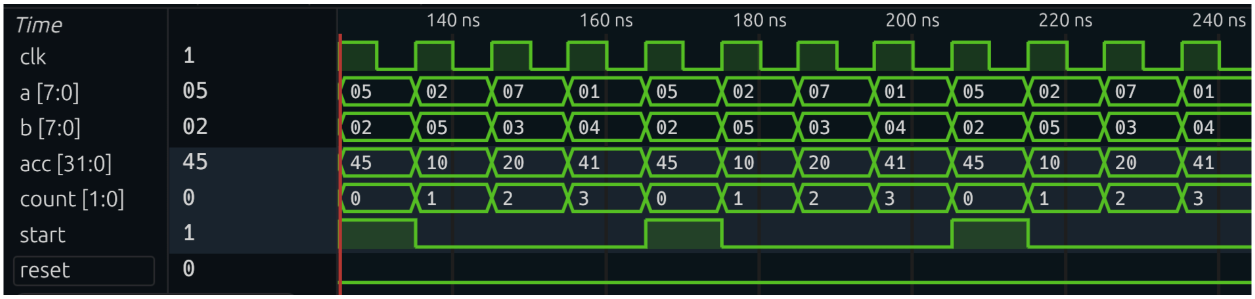
MAC unit block diagram:



Full MAC array block diagram for 4x4 matrices:

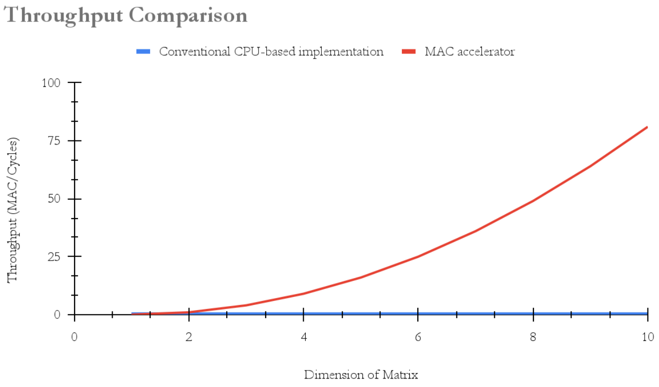


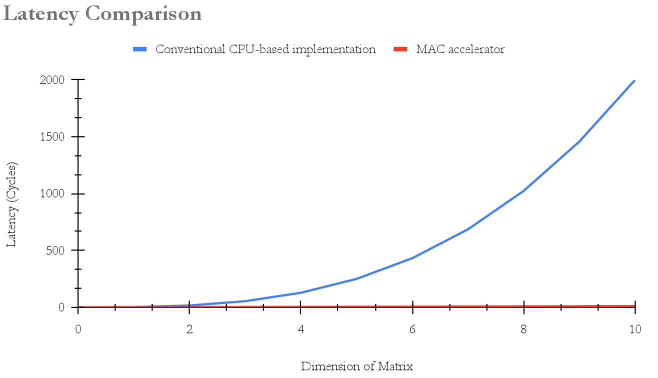
Waveform for one MAC unit:



Performance Metrics:

	Conventional CPU Implementation	**MAC Accelerator**
Throughput	0. 5 MAC/cycle	n^2 MAC/cycle
Latency	$2n^3$ Cycles	n Cycles





The MAC accelerator achieved its performance gains with only a minimal increase in hardware cost because it only adds a small amount of extra hardware.

Conclusion:

This project highlights the promising capabilities of parallel-processing systems for matrix multiplication demanding tasks like AI. Under evaluation, the MAC accelerator I designed dramatically outperformed the conventional CPU implementations in all of the metrics and does so at an affordable cost. Whereas, a conventional single-pipeline CPU performed only 0.5 MAC operations per cycle and requires $2n^3$ cycles to complete an $n \times n$ matrix multiplication, the MAC accelerator achieves n^2 MAC operations per cycle with a latency of only n cycles. These numbers describe the sheer improvement in throughput and substantial reductions in latency as matrix size increases. For tasks like matrix multiplication, the driving factor for efficiency was the amount of tasks done in parallel rather than the raw speed of computation. This project also provided hands-on experience with digital design, parallel, architectures, and Verilog Implementation.