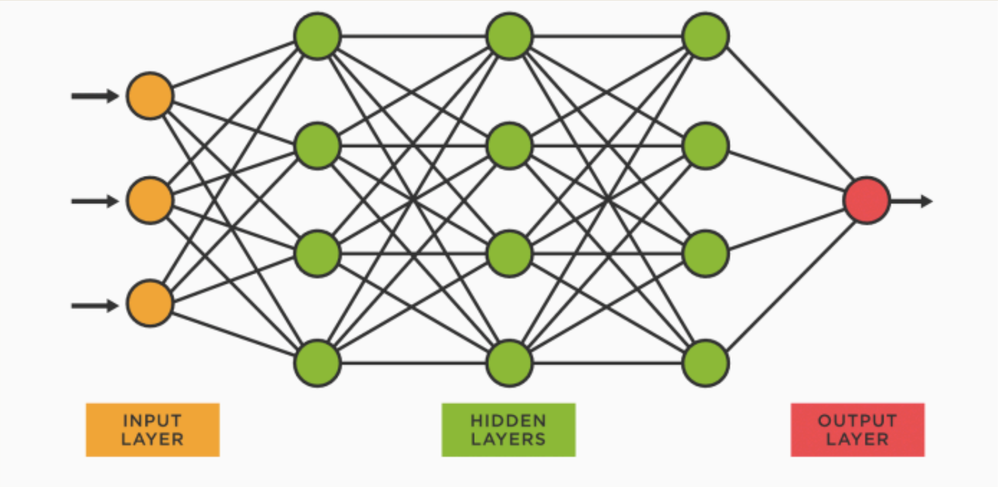
Senior Design Project I
2025 Spring Midterm

Matrix Multiplication Module for a Neural Processing Unit

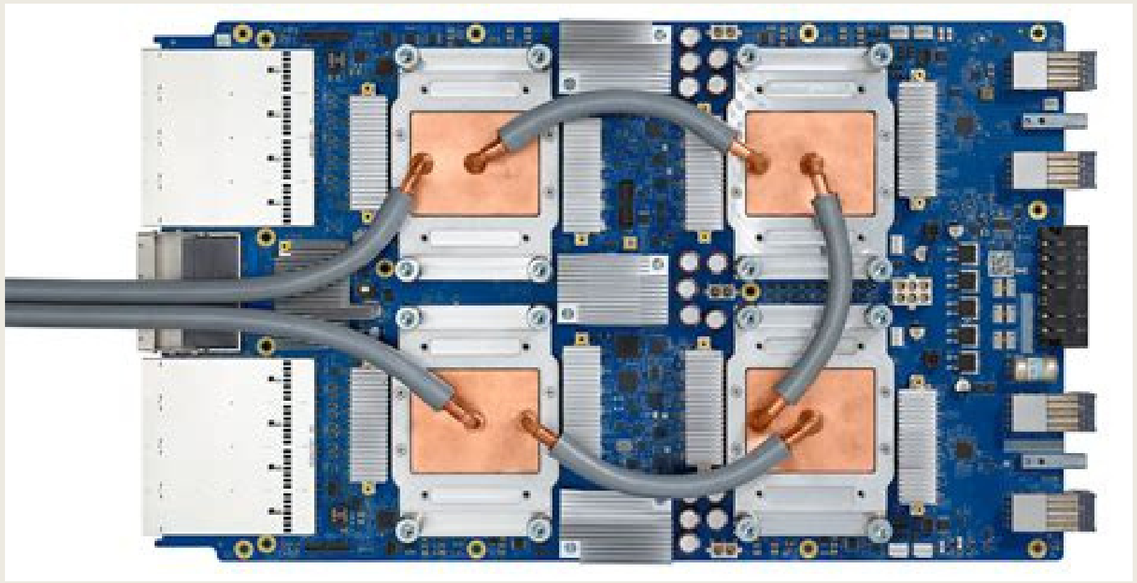
Anastasia Frattarole
Computer Engineering



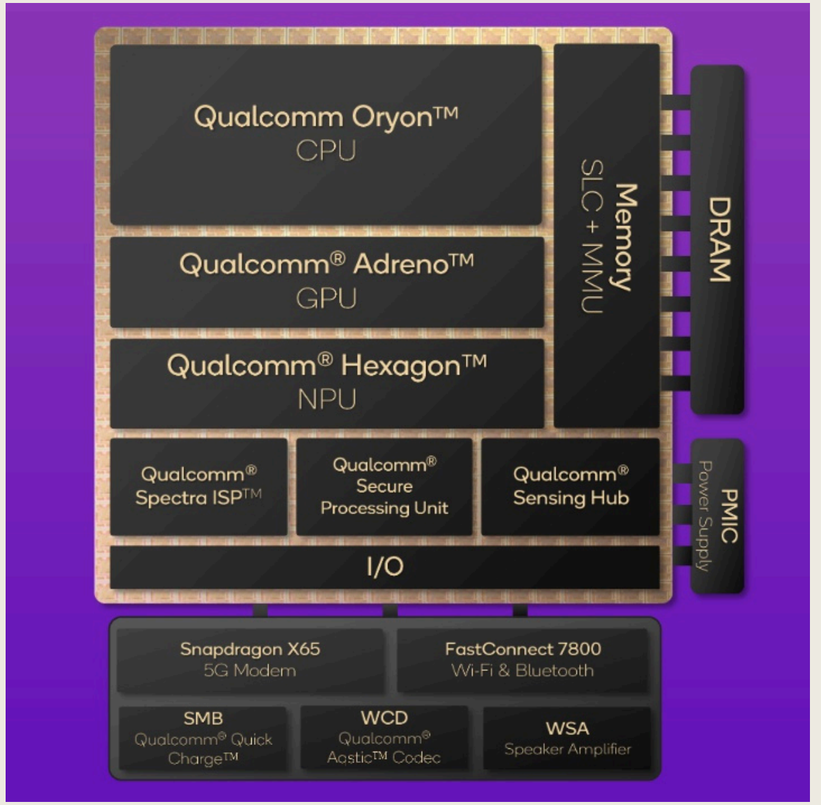
Introduction

- Neural Processing Unit (NPU) aka AI Accelerator
- Specialized processors made specifically for neural network computing
- Each layer uses data from the previous layer to calculate data for the next layer
- Advantages: Power efficient, allows for more parallel processing and hardware acceleration
- Not very well known but becoming more prevalent
- There are several different NPUs out there but they are (usually) not separate from other components

GOOGLE TENSOR PROCESSING UNIT



QUALCOMM HEXAGON NPU



Proposed Solution

- The part of the NPU that a neural network uses to calculate the data needed to make decisions
- Inspired by Google’s TPU
- How it works:
  - Takes two 2x2 16-bit floating point matrices as inputs from memory: a bias matrix and a weight matrix
  - Multiplies these matrices together and produces the sums of a 2x2 16-bit floating point output matrix
  - These sums are put into an activation function module that scales them down to a reasonable size
  - The output matrix is then stored into memory

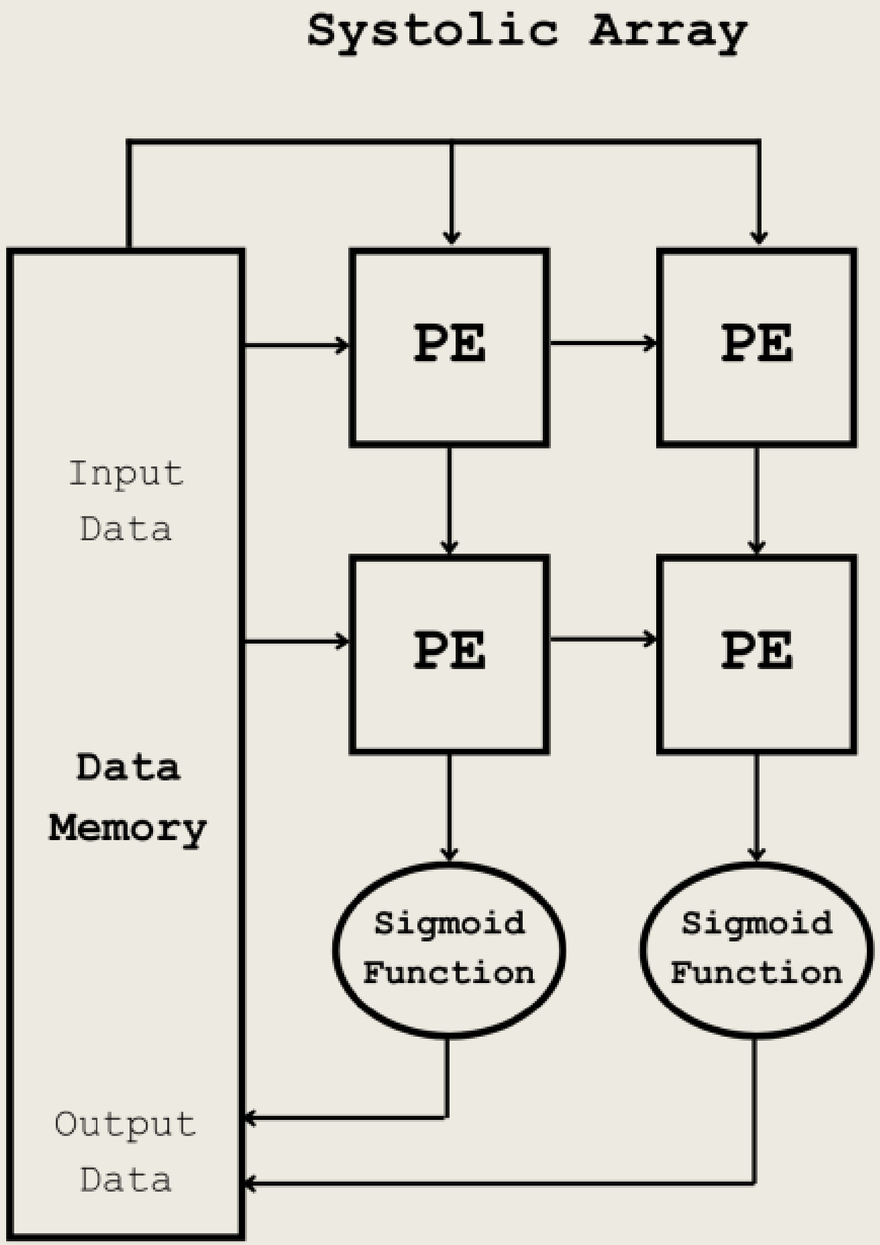
Proposed Solution

- Advantages
  - Uses 16-bit values which allows it to calculate faster
  - Performs multiplication in parallel
  - It is simple, which speeds up computation
- Disadvantages
  - Using 16-bit instead of 32-bit or 64-bit may result in a loss of precision
  - Not a complete NPU
  - It is simple, so it cannot perform anything beyond matrix multiplication

Design Overview

Tools and Standards Used

- Hardware:
  - Altera DE2 FPGA
- Software:
  - Quartus II 13.0sp1
  - ModelSim - Intel FPGA Starter Edition 10.5b
- Programming Languages:
  - Verilog
  - SystemVerilog
- Standards:
  - IEEE 754 (Floating Point Format)
  - IEEE 1364 (Verilog)
  - IEEE 1800 (SystemVerilog)

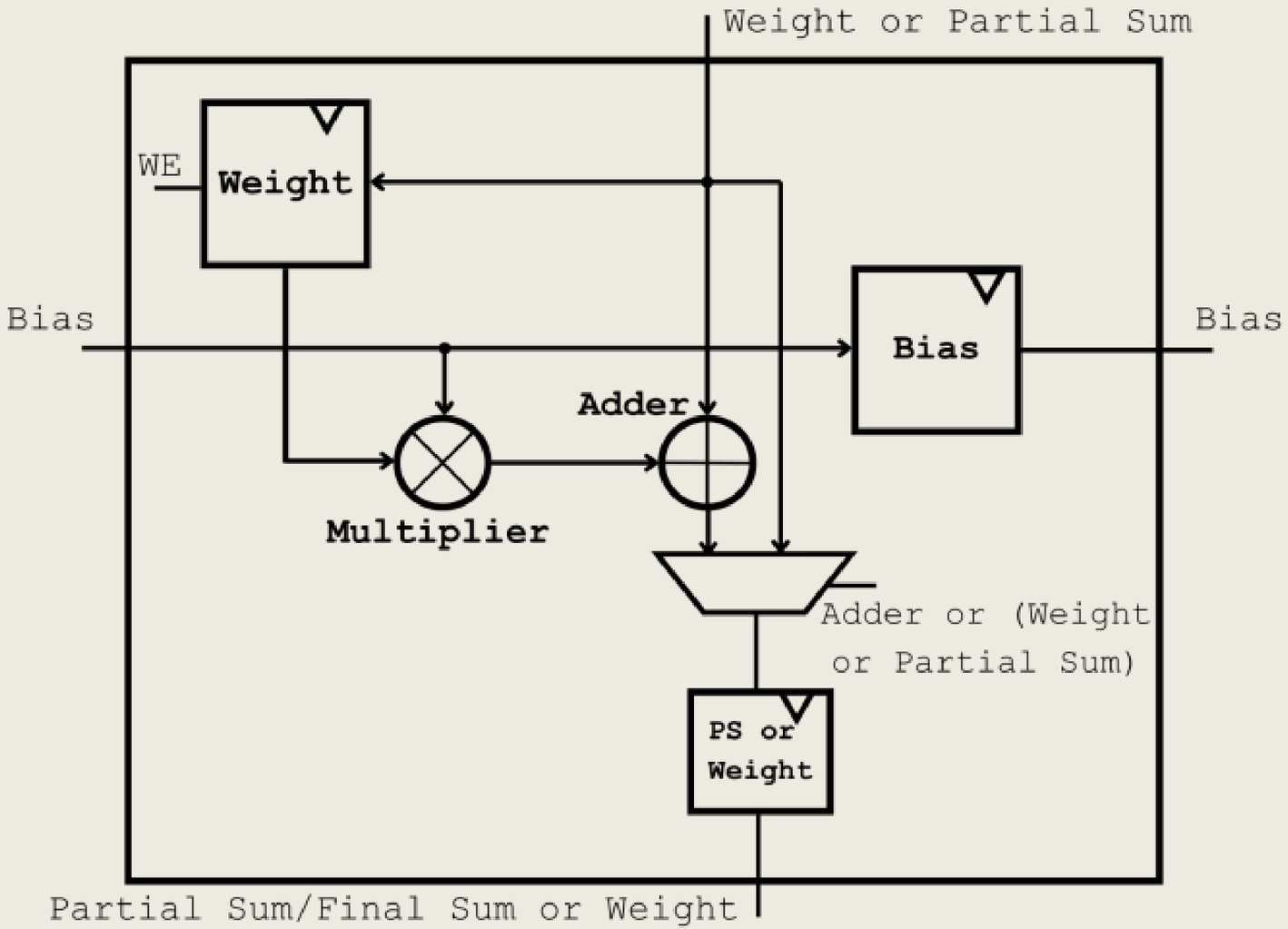


- Inputs:
  - Weight matrix
  - Bias matrix
- Outputs:
  - Output matrix

Systolic Array

- Systolic: Pertaining to a systole or heart contraction
- Weight Stationary
- 16x128 bit memory
- Bias and weight matrices are loaded into the processing elements
- The sums are outputted to the activation functions, which scale the values
- After scaling, the values are stored into memory

Processing Element (PE)



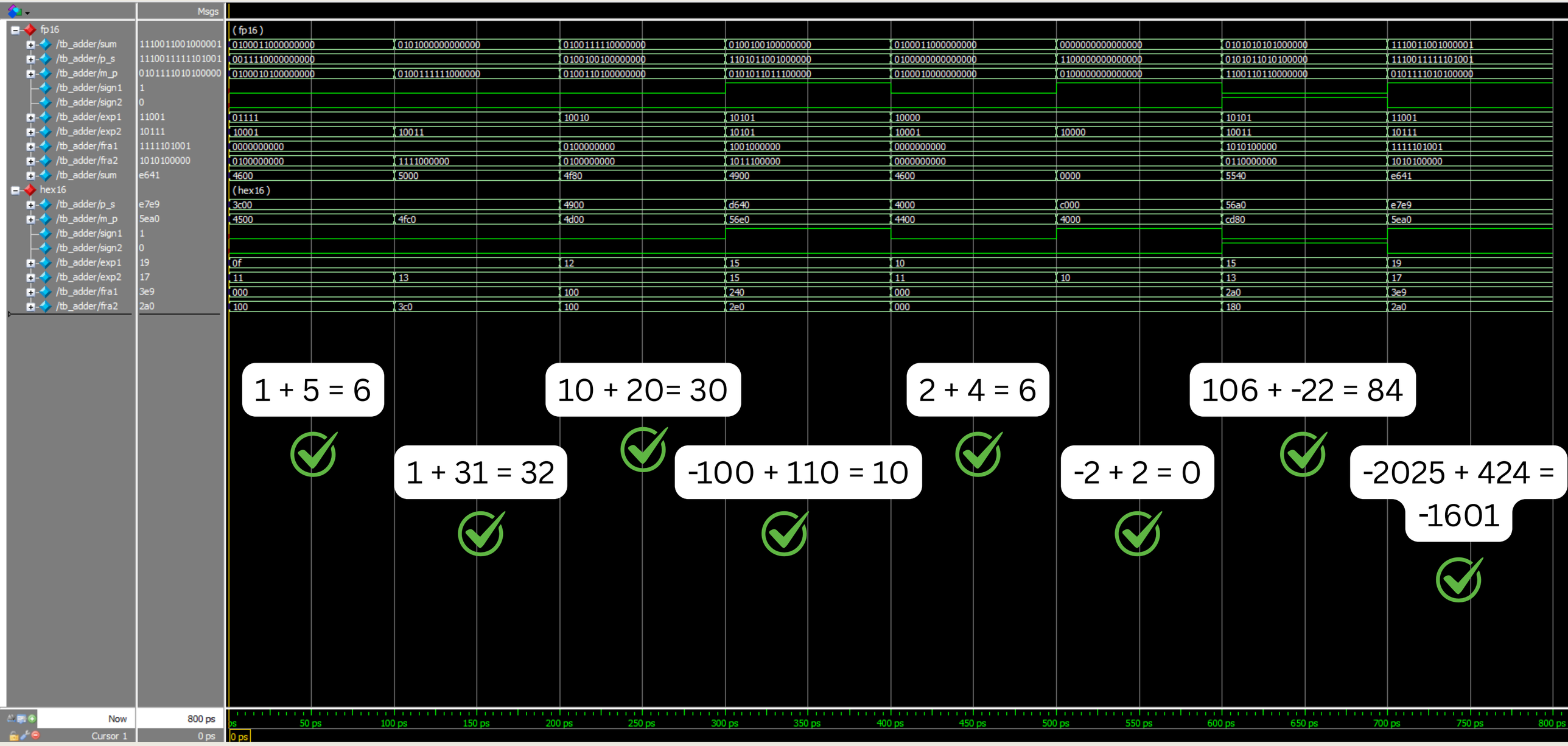
- Inputs:
  - Bias
  - Weight or partial sum
  - Write enable (WE) for weight register
  - Control for multiplexer
- Outputs:
  - Bias (not for PEs in the last column)
  - Updated partial sum or final sum

Processing Element (PE)

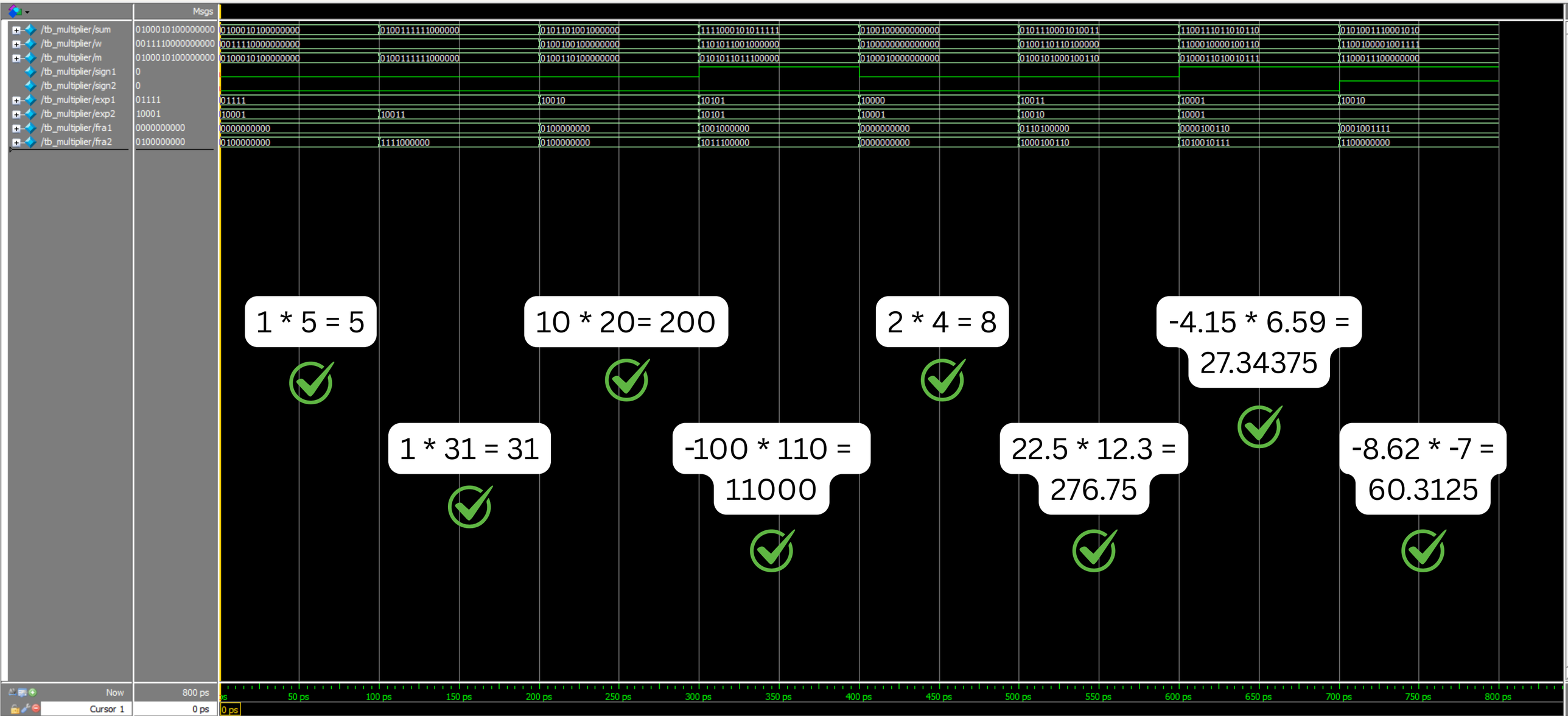
- Weight is loaded into PE before calculations begin
- Multiplies the weight and bias, then adds the product and the partial sum from the PE above it
- Outputs depend on what stage the module is in the calculation
- Can process positive and negative numbers
- Has some exception handling

Simulation Results

Adder Simulation Results



Multiplier Simulation Results



Team Member Responsibilties

Team Member	Contributions and Implemented Functions
Anastasia	• Systolic array design • Processing element design
Anastasia	• Programmed all modules using Verilog or SystemVerilog • Debugged and tested modules
Anastasia	• Made presentation • Researched and verified project information

Conclusions

- Progress so far:
  - Adder and multiplier that perform matrix addition and multiplication with 16-bit positive and negative floating point numbers
- Next steps:
  - Complete PE and Systolic Array modules
  - Implement sigmoid function modules
  - Exception handling (overflow, underflow, etc)
  - Expand to a 4x4 systolic array
- If there is time:
  - Change the adder and multiplier to an ALU that can perform matrix addition, subtraction, multiplication, and division

Acknowledgements

- Dr. Zhu - helped me come up with the idea for the project
- Sam - helped me with Canva to make the Systolic Array and PE visual designs
- My family and friends

References

[1] “NPU vs. GPU: What’s the Difference?,” Micro Center, Oct. 16, 2024. https://www.microcenter.com/site/mc-news/article/npu-vs-gpu.aspx

[2] IEEE SA, “IEEE Standards Association,” IEEE Standards Association. https://standards.ieee.org/ieee/754/6210/

[3] J.-Y. Kim, “Hardware accelerator systems for artificial intelligence and machine learning,” in FPGA based neural network accelerators, S. Kim and G. C. Deka, Eds. Elsevier, 2021.

[4] N. Kung, “Why systolic architectures?,” Computer, vol. 15, no. 1, pp. 37–46, Jan. 1982, doi: 10.1109/mc.1982.1653825.

[5] A. Mohan, “Understanding matrix multiplication on a Weight-Stationary Systolic architecture | Telesens,” Telesens, Feb. 19, 2019. https://www.telesens.co/2018/07/30/systolic-architectures/

[6] T. Raja, “Systolic Array Data Flows for Efficient Matrix Multiplication in Deep Neural Networks,” arXiv, Art. no. arXiv:2410.22595, [Online]. Available: https://arxiv.org/abs/2410.22595

[7] V. Sze, Y.-H. Chen, T.-J. Yang, and J. S. Emer, “Efficient Processing of deep Neural Networks: A tutorial and survey,” Proceedings of the IEEE, vol. 105, no. 12, pp. 2295 2329, Nov. 2017, doi: 10.1109/jproc.2017.2761740.

Images:

https://medium.com/@decoded_cipher/tensor-processing-units-both-history-and-applications-b3479d92a61d

https://www.notebookcheck.net/Qualcomm-releases-official-Snapdragon-X-Plus-and-Snapdragon-X-Elite-benchmarks-for-45-TOPS-Hexagon-NPU.841811.0.html

https://www.marktechpost.com/2022/09/23/top-neural-network-architectures-for-machine-learning-researchers/

Thank you!