IOWA STATE UNIVERSITY
Department of Computer Engineering

Honors Capstone Project: Fall 2020
Honors Poster Session
December 7th, 2020

Matthew Dwyer

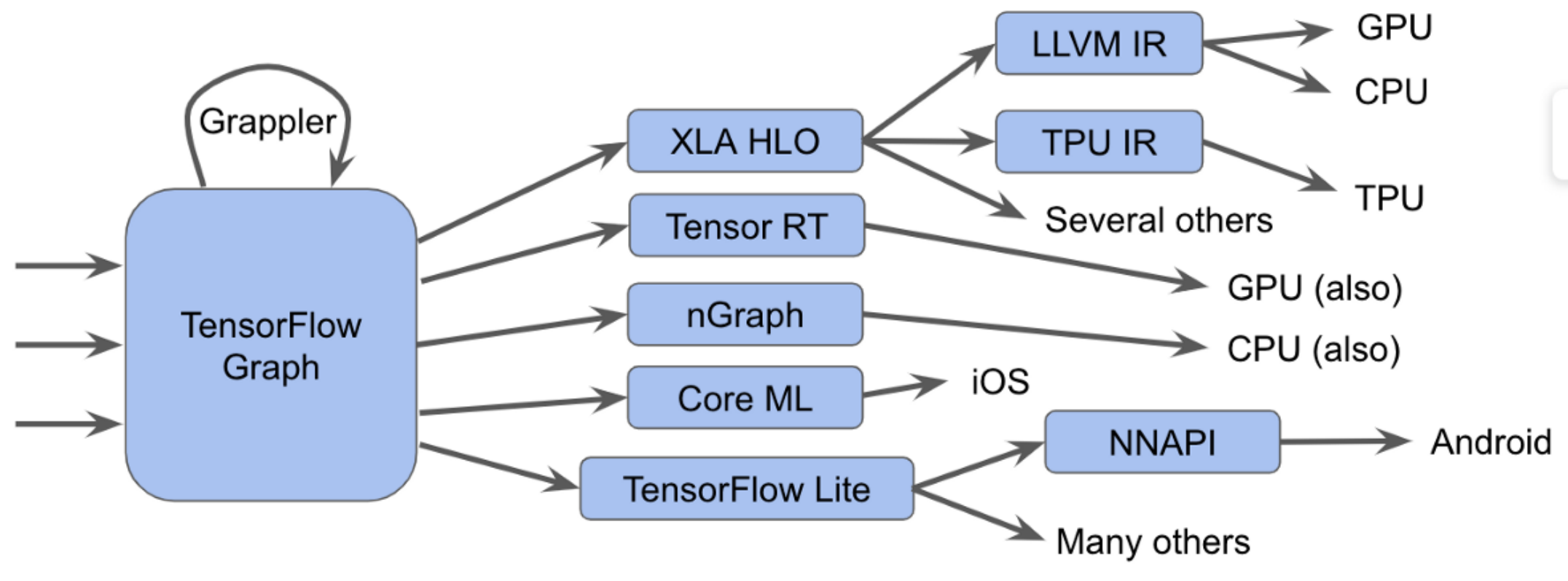
Advised by: Dr. Henry John Duwe

Machine Learning Accelerated Hardware Designs & Software Implementations

Project Objectives

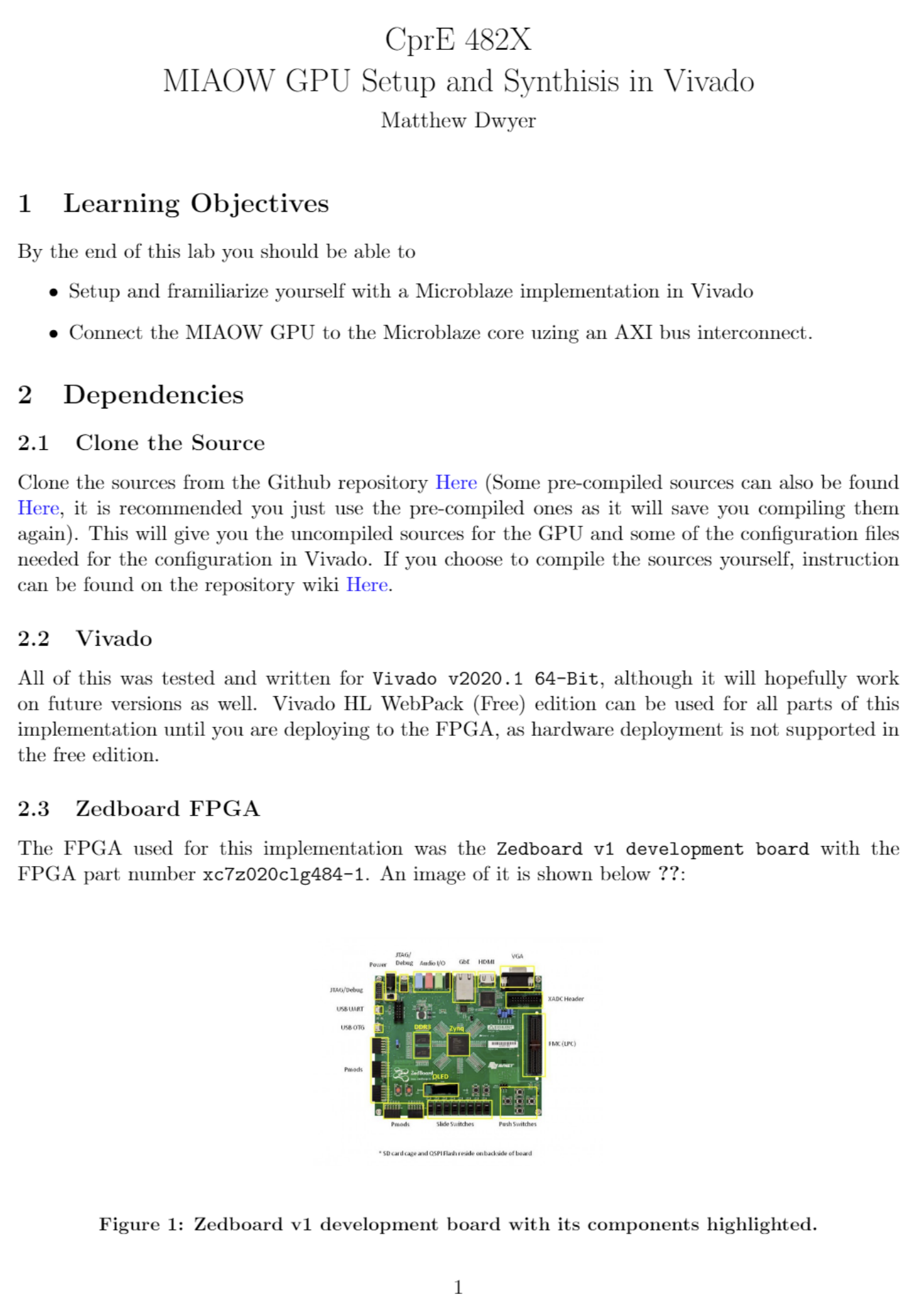
This project focused on the following objectives to guide deliverable content for the CPR E 482X course offering:

- Develop labs that are capable of being utilized as effective teaching tools in a classroom setting for the comprehension of machine learning and hardware acceleration opportunities and designs:
  - Create effective lab environments to aid students in utilizing complex machine learning tools and libraries.
  - Train and document machine learning models (ResNet) on a variety of datasets (ImageNet, MNIST).
- Identify and explore available FPGA-based machine learning accelerated architectures (MIAOW, VTA).



Tensorflow model graph backend compilation, optimization, and IR flow for various hardware platforms.

Nvidia Nsight GPU kernel performance analysis tool.



CprE 482X
MIAOW GPU Setup and Synthisis in Vivado
Matthew Dwyer

1 Learning Objectives

By the end of this lab you should be able to

- Setup and framiliarize yourself with a Microblaze implementation in Vivado
- Connect the MIAOW GPU to the Microblaze core uzing an AXI bus interconnect.

2 Dependencies

2.1 Clone the Source

Clone the sources from the Github repository Here (Some pre-compiled sources can also be found Here, it is recommended you just use the pre-compiled ones as it will save you compiling them again). This will give you the uncompiled sources for the GPU and some of the configuration files needed for the configuration in Vivado. If you choose to compile the sources yourself, instruction can be found on the repository wiki Here.

2.2 Vivado

All of this was tested and written for `Vivado v2020.1 64-Bit`, although it will hopefully work on future versions as well. Vivado HL WebPack (Free) edition can be used for all parts of this implementation until you are deploying to the FPGA, as hardware deployment is not supported in the free edition.

2.3 Zedboard FPGA

The FPGA used for this implementation was the `Zedboard v1 development board` with the FPGA part number `xc7z020clg484-1`. An image of it is shown below ??:

Figure 1: Zedboard v1 development board with its components highlighted.

1

Lab documentation for re-implementation of the MIAOW GPGPU in Vivado design software for the Zedboard FPGA.

Methods

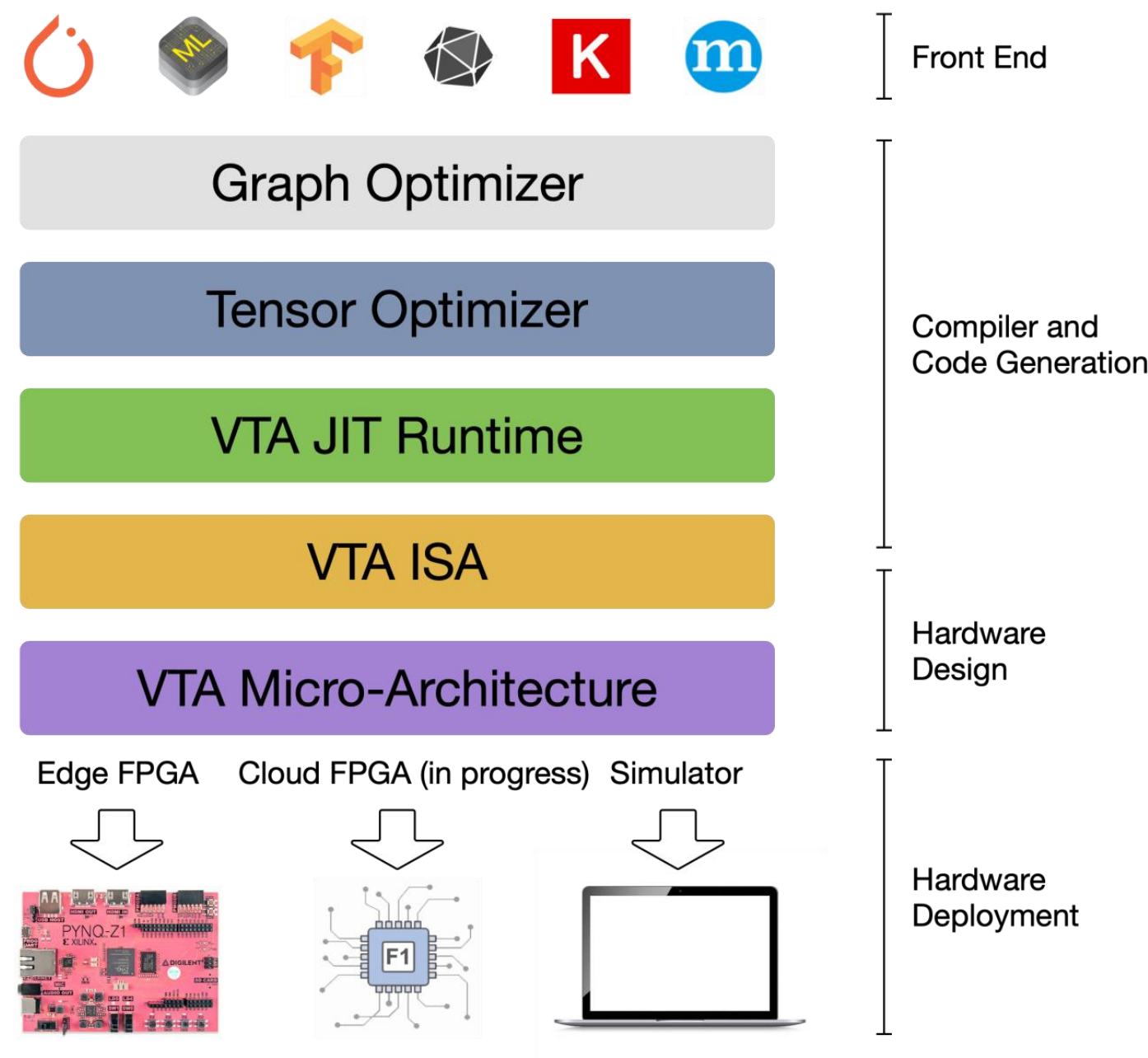
The following steps were taken to realize the aforementioned goals of this project, each designed to build off the previous:

- Evaluated other popular architecture designs of research machine learning accelerators:
  - Apache Versatile Tensor Accelerator (VTA) [Open Source]
  - Many-core Integrated Accelerator of Waterdeep/Wisconsin (MIAOW) [Open-Source]
- Evaluated optimization steps and low-level IRs of machine learning frameworks.
- Implemented Python tooling for lab creation with Jupyter notebooks implementing Tensorflow, TVM on toy Machine Learning applications.
- Documenting re-implementation of architectural designs in system design software (Vivado) for Zedboard FPGA development boards.

Results

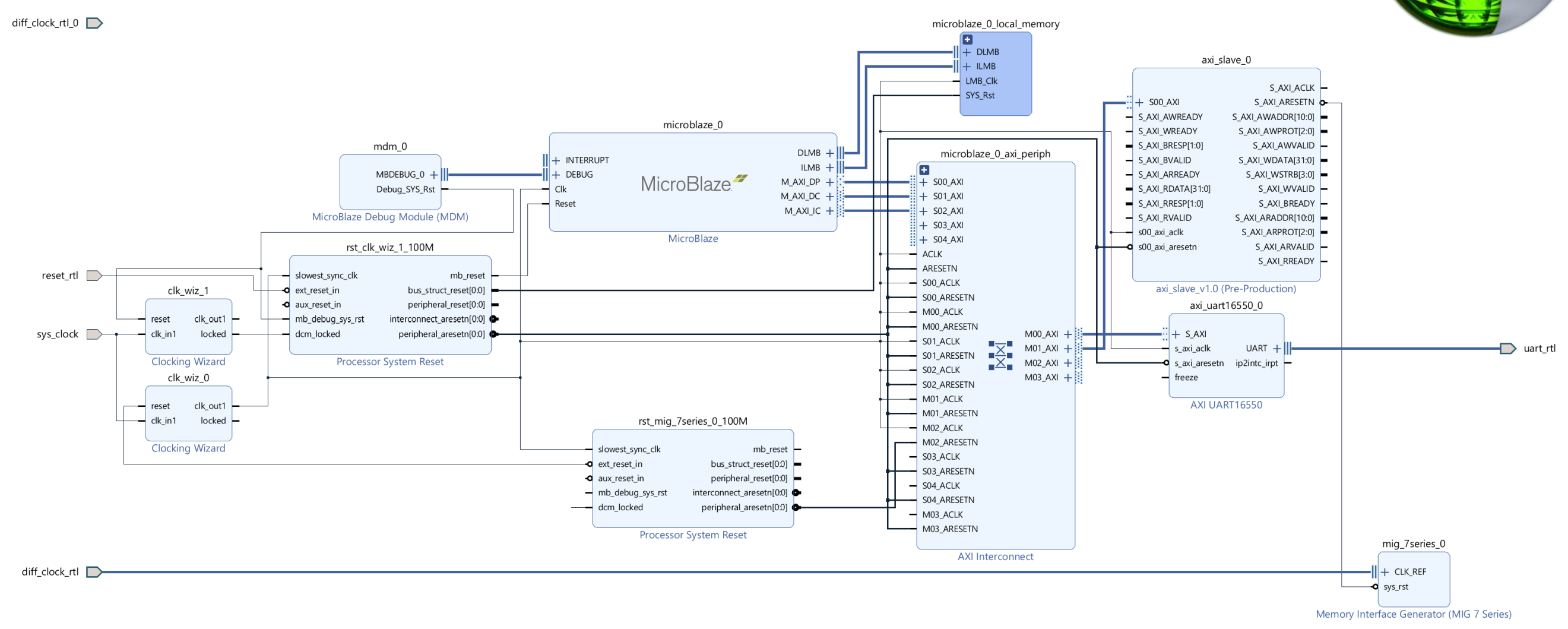
The deliverables of this project were created and evaluated on their ability to be used as laboratory exercises and tools by students in future offerings of the CPR E 482X course and are as follows:

- Documentation explaining project synthesis in Vivado design software of the open-source MIAOW GPGPU accelerator.
- Laboratory tools written in Python utilizing the Jupyter notebooks, the Tensorflow framework, Anaconda environment manager, and TVM compilation framework.
- Research into current machine learning compilation frameworks and their interconnections to hardware and software acceleration. Rejected hardware setups focusing on machine learning performance rather than embedded design:
  - Modifying the XLA backend of Tensorflow for custom function implementations.
  - Nvidia CUDA GPGPU programming for TensorCore ML acceleration.



Apache VTA hardware acceleration and compilation structure.

Tools & Technologies



MIOAW GPU FPGA implementation in Vivado design software. Most logic is devoted to the MicroBlaze core which controls the MIAOW GPU over a master-slave AXI4 bus interconnect.

Outcomes

The resulting work would go on to be utilized in the following ways:

- Hardware acceleration design research and exploration enhanced course topics and areas of study.
- Lab environment tooling and utilities utilized on the Fall 2020 course offering of CPR E 482X.
- MIAOW GPGPU implementation documentation and examples plan to be utilized on future CPR E 482X offerings to enhance lab exercise's connection to course content.

Work done as part of the preparation of course and lab materials for CPR E 482X: Hardware Design for Machine Learning.