Imperial College London

MEng Individual Project

Imperial College London

Department of Computing

Reconfigurable Acceleration of Transformer Neural Networks with Meta-Programming Strategies for Particle Collisions Experiments

Author:
Filip Wojcicki

Supervisor:
Prof. Wayne Luk

Second Marker:
Dr. TODO

January 22, 2022

Contents

Chapter 1

Introduction

1.1 Motivation

Physics experiments are crucial Big data is crucial for Physics LHC is currently bottleneck at real time speed for detectors Transformer Neural Networks are great

Powerful hardware is king FPGA are great for NN Metaprogramming allows for optimizations and customisability

1.2 Objectives and Challenges

Current architecture at LHC is too slow -> make it quick enough for real time processing HLS is difficult, so coding hardware in Python is desired -> make it easy for engineers and physicists to design systems FPGA are very hard-coded -> make the code deployable on any platform with optimal settings automatically

1.3 Contributions

hls4ml

Chapter 2

Background

training - inference - tensor - particle Collisions jets latency resource usepackage throughput pipelining HLS hardware design (parallel vs serial etc)

Chapter 3

Project Plan

The aims of the project cover a wide range of challenges that form subsequent steps of accelerating neural networks while raising the abstraction layers and reducing domain-specific knowledge requirements. This naturally divides the work into smaller objectives that are described in details in the following paragraphs.

Firstly, the existing transformer neural network architecture has to be redesigned to accommodate for easier adaptation to non general-purpose hardware. This comprises of splitting layers into more basic components that are easier to map to hardware and abstract about as well as introducing hooks that collect different information during training and inference passes (e.g. running mean and variance for normalization layers, tensor sizes and values). At this phase some of the design choices are highlighted for further inspection where simplification or improvements can be made to greatly reduce the complexity and resource usage without crippling performance.

With the adapted software implementation, the next step involves recreating the architecture in HLS. Building the initial prototype tackles the difficulties related to the underlying differences between software and hardware development and results in an accurate, yet not optimal design. From there, an iterative process begins with acceleration hypothesis firstly tested in the original software model to ensure satisfactory accuracy and then getting expressed in HLS to quantitatively measure the latency and throughput differences. That is expected to yield a highly performant solution to the initial problem that is tailored to the specific FPGA constraints.

In order to overcome the innate limitations of "hand-tuning" a solution to a problem that varies both in time and between applications, the final step of the project relies on meta-programming strategies that automatically adapt the solution according to users' criteria, available platforms and overall experiment's aim. The list of approaches that can be taken here is nearly endless, however two key areas have be designated - adjusting the model according to the existing hardware to exploit its strengths as well as allowing for more abstract representation of an architecture in a well-known machine learning framework.

As previously mentioned, some of the initially planned ideas have already been implemented. The distinction between these and a more detailed look at the specific project tasks can be seen in figure 3.1.

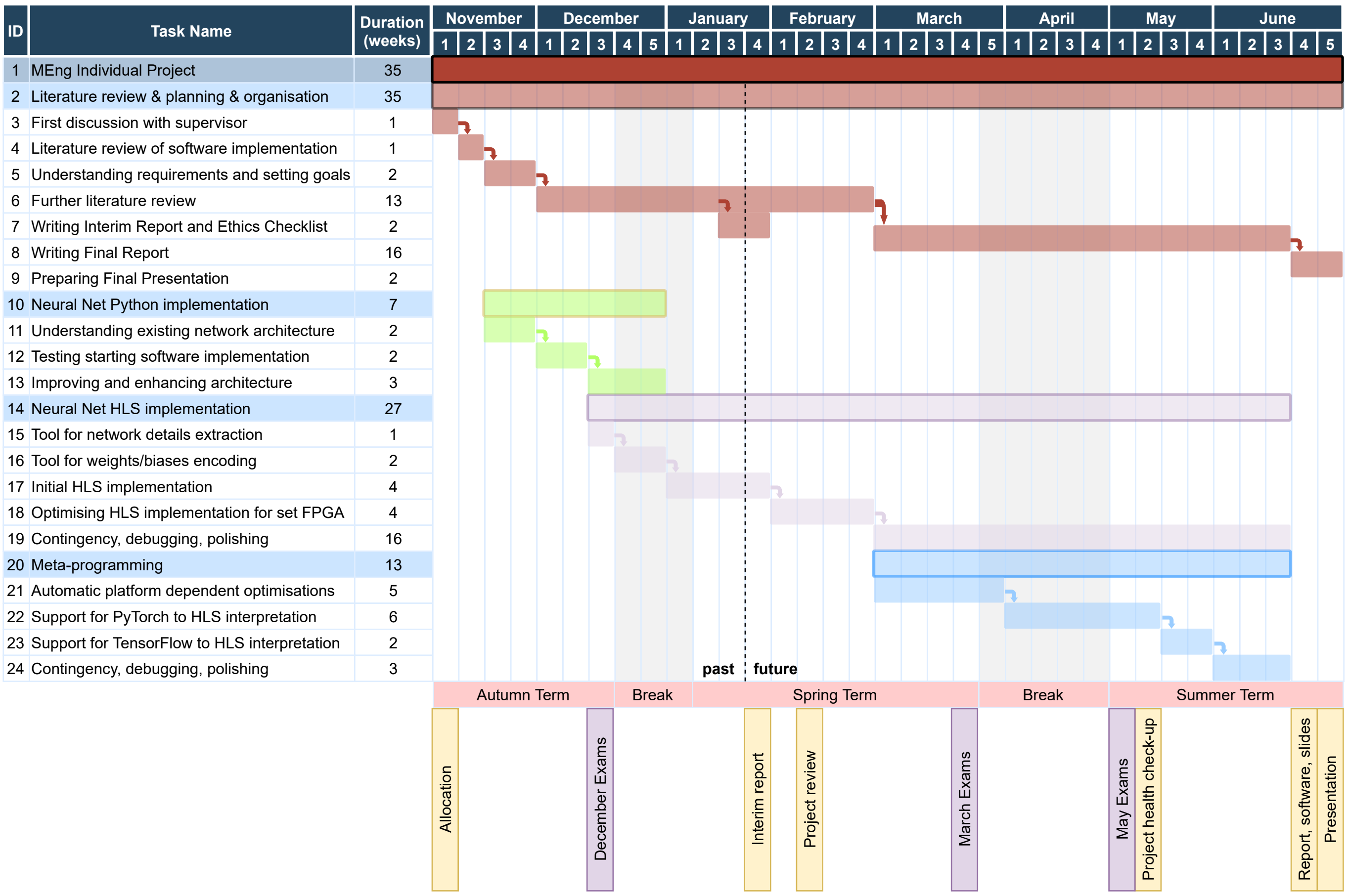

Figure 3.1: Project's Gantt chart representing initial plan of the work, past schedule has been updated to match ongoing progress accordingly

Chapter 4

Implementation

Chapter 5

Evaluation Plan

Chapter 6

Ethical Issues

Table 6.1: Overview of potential categorized ethical issues with an indication of their applicability

	Involvement of...	Exists?
Humans	human participants	No
Personal data	personal data collection and/or processing	No
	collection and/or processing of sensitive personal data	No
	processing of genetic information	No
	tracking or observation of participants	No
	further processing of previously collected personal data	No
Animals	animals	No
Developing countries	developing countries	No
	low and/or lower-middle income countries	No
	putting the individuals taking part in the project at risk	No
Environment	elements that may cause harm to the environment, animals or plants	*
	elements that may cause harm to humans	*
Dual use	potential for military applications	No
	strictly civilian application focus	Yes
	goods or information requiring export licenses	No
	affection of current standards in military ethics	No
Misuse	potential for malevolent/criminal/terrorist abuse	*
	information on/or the use of sensitive materials and explosives	No
	technologies that could negatively impact human rights standards	*
Legal	software for which there are copyright licensing implications	No
	information for which there are data protection or other legal implications	No
Other	anyother ethics issues that should be taken into consideration	No

Bibliography