
Contents

List of Figures	xvii
List of Tables	xxvii
Glossary	xxxix
Acronyms	xxxv
Research Context and Resources	xxxix
Publications and Original Work	xli
1 Introduction	1
2 Motivation	3
2.1 Real-Time Analysis	3
2.2 Real-Time Analysis in High-Energy Physics	7
2.3 RTA using ML on Heterogeneous Architectures	12
I Background	15
3 Physics Background	17
3.1 Accelerator Physics	17
3.2 The Standard Model of Particle Physics	23
3.3 Open Questions	24
3.4 Heavy Flavor Physics	26
4 Machine Learning Background	29
4.1 Machine Learning	29
4.2 Deep Learning	37
4.3 Convolutional Neural Networks	43
4.4 Graph Neural Networks	45
4.5 Quantization	49
5 High Performance Computing	53
5.1 Parallelism	53
5.2 From Video Games to the GPU	58
5.3 CUDA Programming Model	59

5.4	Programmable Logic	63
5.5	Field-Programmable Gate Arrays	65
6	The LHCb Experiment at CERN	69
6.1	The Large Hadron Collider at CERN	69
6.2	LHCb Detector Overview	73
6.3	Vertex Locator	75
6.4	Online System and Data Acquisition	77
6.5	Software Framework	79
6.6	Simulation	80
6.7	LHCb Trigger System	81
7	Track Reconstruction	91
7.1	Track Reconstruction	91
7.2	Track Reconstruction at LHCb	92
II	Main Results	103
8	ETX4VELO: Tracking with GNNs at LHCb	105
8.1	Early Version of ETX4VELO	106
8.2	Reconstruction of Electrons	111
8.3	The ETX4VELO Pipeline	116
8.3.1	Datasets	117
8.3.2	Hit Embedding and Rough Graph Construction	118
8.3.3	Graph Neural Network and Classifiers	119
8.3.4	Triplet Building	122
8.3.5	Track Building	122
8.3.6	Training Process	123
8.4	Physics Performance	123
9	Accelerating ETX4VELO on GPU	131
9.1	Development	132
9.2	The ETX4VELO Pipeline on GPU	137
9.2.1	Structure of Data in Allen	137
9.2.2	Network Inference	139
9.2.3	k-NN Implementation	142
9.2.4	WCC Implementation	142
9.2.5	Quantization	143
9.2.6	Physics Performance	145
9.3	Computational Performance	145
9.4	Throughput Scaling Comparison	148

10 Accelerating ETX4VELO on FPGA	153
10.1 Implementation of the Embedding	155
10.1.1 PYNQ Framework	156
10.1.2 PYNQ-Z2 Development Board	156
10.1.3 Workflow	159
10.1.4 Evaluation of Precision Loss	161
10.2 Latency Comparison of ML Model Inference	162
10.3 Throughput Comparison of ML Model Inference	163
10.4 Purchase vs. Operating Cost	168
11 Conclusion and Outlook	171
 III Appendices	 173
A Notations, Units and Physical Constants	175
A.1 Notations	175
A.2 Units and Abbreviations	176
A.3 Physical Constants	176
B Early ETX4VELO Development	177
C Further Resources	181
Bibliography	183