Table of Contents

1 Introduction 1

2 Background & Related Work 3

2.1 Introduction to NVIDIA Ampere Architecture 3

2.2 Transformers 4

2.2.1 T5 6

2.2.2 NLLB 7

2.2.3 NLLB-MoE & Switch Transformers 7

2.3 Performance Evaluation on MT models 8

2.3.1 Quality Metrics 8

2.3.2 Model Inference Time 11

2.4 Disk offloading 12

2.4.1 Related work: Deepspeed 12

2.4.2 Archer 13

3 Benchmark Design and Implementation 15

3.1 BenchmarkingSuite 15

3.2 Configuration Setup 17

3.2.1 ExperimentConfiguration 17

3.3 Dataset Class 17

3.3.1 Archer 18

3.4 Analytics 18

3.4.1 Quality Metrics 19

3.4.2 Efficiency Metrics 20

3.4.3 Profiling 20

3.5 Implementation 20

4 Experiments **22**
4.1 Datasets & Models 22
4.2 Hyperparameters 25
4.3 Methodology 27
4.4 Results & Analysis 29
4.4.1 Dense Models 29
4.4.2 Sparse Models 34

5 Conclusions **39**

Bibliography **41**

6 Appendix **47**
6.1 Single sentence dataset from FLORES-200 for Examining Efficiency of Models 47
6.2 Single sentence dataset from WMT14 for Examining Efficiency of Models 47
6.3 Training hyperparameters of the SwitchTransformer Model Fine-Tuned 48