RWTH Aachen University
Department of Computer Science

Design and Performance Engineering of GPU-Accelerated Tensor Network Algorithms for Large-Scale Scientific Simulations

Master Thesis

submitted by

Daniel Sinkin
Matriculation Number: 367316

First Examiner:	Prof. Dr. TODO
Second Examiner:	Prof. Dr. TODO
Advisors:	Dr. Edoardo Di Napoli
External Institution:	Jülich Supercomputing Centre (JSC) Forschungszentrum Jülich

Aachen, February 17, 2026

Abstract

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Acknowledgements

TODO

Contents

List of Figures

List of Tables

Listings

1 Introduction

1.1 Motivation

1.2 Problem Statement

1.3 Contributions

1.4 Outline

The remainder of this thesis is structured as follows. ?? introduces ...?? presents ...?? details ...?? evaluates ...?? summarises ...

2 Background

2.1 Tensor Networks

2.1.1 Tensor Notation and Diagrams

2.1.2 Tensor Contraction

2.1.3 Tensor Network Structures

2.2 GPU Architecture

2.2.1 Streaming Multiprocessor and Warp Execution

2.2.2 Memory Hierarchy

2.2.3 NVIDIA A100 Ampere Architecture

2.3 CUDA Programming Model

2.3.1 Thread Hierarchy and Kernel Launch

2.3.2 Shared Memory and Synchronisation

2.3.3 Memory Coalescing and Bank Conflicts

2.3.4 Performance Profiling with Nsight Compute

2.4 Related Work

2.4.1 cuBLAS and cuTENSOR

2.4.2 ChASE Eigensolver

2.4.3 Existing GPU Tensor Network Implementations

3 Design and Methodology

3.1 Target Kernels

3.2 Algorithmic Approach

3.3 Data Layout and Memory Strategy

3.4 Baseline Selection

4 Implementation

4.1 Kernel Design

4.2 Shared Memory Tiling

4.3 Occupancy and Launch Configuration

4.4 Integration and Build System

5 Results

5.1 Experimental Setup

5.2 Single-GPU Performance

5.3 Profiling Analysis

5.4 Comparison with cuBLAS and cuTENSOR

5.5 Scaling Behaviour

5.6 Discussion

6 Conclusion

6.1 Summary

6.2 Limitations

6.3 Future Work

A Supplementary Benchmarks

TODO

Declaration of Authorship

I hereby declare that I have created this work completely on my own and used no other sources or tools than the ones listed, and that I have marked any quotes accordingly.

Aachen, February 17, 2026 Daniel Sinkin