
Cade Brown

Agentic performance optimization | Compilers & DSLs | GPU numerical libraries

brown.cade@gmail.com · linkedin.com/in/cade-brown · github.com/cadebrown · cade.io

Summary

Senior engineer working on CUDA math libraries, compiler integration, and agent-driven kernel development, with research in numerical computing and scientific machine learning.

Technical Focus

CUDA, C++, Python, GPU performance engineering, numerical software, compiler and IR development, JIT/LTO, MLIR, LLVM, SIMT programming, NEON, SVE, scientific computing

Experience

Senior CUDA Math Libraries Engineer | NVIDIA

March 2025–Present | Ann Arbor, MI (Remote)

- Developed a framework for AI-agent-driven CUDA kernel generation, profiling, optimization, modification, and autotuning.
- Developed modular, agent-friendly domain-specific languages for algebraic code transformations and code reuse.
- Developed CUDA Tile integration and generalized autotuning with SIMT interoperability, cuTile Python, NVCC, LTO IR, and compiler technologies.
- Built developer productivity tooling including harnesses, CLIs, reusable setups, sandboxing, onboarding guides, and GPU CI monitoring and triage.
- Studied GPU I/O topologies, algorithmic changes, and simulated performance to inform hardware/software co-design.

CUDA Math Library Engineer | NVIDIA

August 2023–March 2025 | Ann Arbor, MI (Remote)

- Optimized CUDA math-library and cuFFT performance through memory-bandwidth work, asynchronous I/O strategies, and code generation.
- Used JIT LTO to generate optimized kernels at runtime for particular user problems while reducing required precompiled kernels.
- Optimized NVPL CPU math libraries on Grace CPUs with NEON and SVE vectorization.

Experience (continued)

Research Assistant (HPC) | Innovative Computing Laboratory, University of Tennessee, Knoxville

August 2019–May 2023 | Knoxville, TN

- Developed GPU compute kernels for MAGMA and SLATE for scientific and machine-learning applications on supercomputers.
- Optimized MAGMA BLAS for AMD GPUs; related work appears in a coauthored IEEE HPEC 2020 paper.
- Designed SABATH, a platform for reproducible scientific machine-learning applications following FAIR principles; related work appears in a coauthored IEEE HPEC 2022 paper.

Tools: CUDA, C/C++, Fortran, Python, MPI, OpenMP, HIP/ROCM, DPC++/oneAPI, PyTorch.

Software Architect | InterGauge

August 2022–April 2023 | Knoxville, TN

- Provided technical leadership for an education-technology startup focused on interactive educational AI assistants and classroom focus metrics.

Machine Learning Intern | NVIDIA

May 2022–September 2022 | Remote

- Researched hardware acceleration for neural-network training and inference through DSLs, intermediate representations, kernel generation, and polyhedral GPU compilation.
- Trained and evaluated models for code generation and optimization-pass selection.

Tools: CUDA, C/C++, Python, LLVM/MLIR, NumPy, PyTorch, TensorFlow, Matplotlib.

Additional Experience

Research Assistant (HCI) | University of Tennessee, Knoxville

August 2021–January 2022 | Knoxville, TN

- Developed Avocat, a terminal tool that interprets error output and suggests shell-command solutions using generative text models and semantic vector search.
- Optimized graph-database queries for analytics over a terabyte-scale source-code dataset.

Tools: Java, Python, REST, ArangoDB.

Machine Learning Consultant | Qardian Labs

January 2020–April 2020 | Knoxville, TN

- Trained and validated a model for heart-disease identification and anomaly detection, with a web interface for risk assessment and further-investigation decisions.

Tools: Python, TensorFlow, Django, Heroku, Google Cloud SDK.

Additional Experience (continued)

Research Intern (HPC) | Oak Ridge National Laboratory

May 2017–August 2018 | Knoxville, TN

- Contributed with Thomas Hill to SimpleSummit, a portable NVIDIA Jetson cluster used to explain distributed computing near Summit at ORNL.
- Developed FractalExplorer, a distributed GPU-accelerated interactive visualization for Mandelbrot and other fractals on SimpleSummit.
- Implemented and ran benchmarks to measure Lustre parallel-filesystem scalability on the Titan supercomputer.
- Assisted with fabrication of the SimpleSummit cluster's 3D-printed enclosure for eight NVIDIA Jetson boards.

Tools: C/C++, CUDA, Python, MPI, SDL, Blender.

Software Consultant | Agilaire LLC

- Designed and implemented pilog, a hardware/software air-quality data logger with sensor collection, time-series storage, visualization, and programmatic access.

Tools: Python, Django, hardware, HTML, JavaScript, CSS.

Publications

Piotr Luszczek and Cade Brown. [Surrogate ML/AI Model Benchmarking for FAIR Principles' Conformance](#). IEEE HPEC, 2022.

Cade Brown, Ahmad Abdelfattah, Stanimire Tomov, and Jack Dongarra. [Design, Optimization, and Benchmarking of Dense Linear Algebra Algorithms on AMD GPUs](#). IEEE HPEC, 2020.

Talks & Presentations

SABATH: ML Surrogate Platform

MLCommons Science | 2022

- Presented SABATH's approach to reproducibility in scientific machine-learning models.

Challenge #6: Sentinel-3 Satellite Dataset

Smoky Mountain Conference, ORNL | 2022

- Introduced the SMCEFR dataset and data challenge for participating researchers.

MAGMA to hipMAGMA: Adding HIP Compatibility to a Scientific C++ Library

ICL Annual Conference | 2018

- Demonstrated HIP portability for CUDA scientific libraries targeting AMD GPUs.

Research Service

Dataset Sponsor | Smoky Mountain Conference, ORNL | 2022

Contributed the SMCEFR Sentinel-3 dataset with Piotr Luszczek for the conference's research data challenge.

Research

Cade Brown. OEIS A267263, 2016.

SMCEFR satellite-data contribution with Piotr Luszczek: a 1024×1024 RGB subset derived from Sentinel-3 OLCI EFR data; Cade Brown is identified as corresponding author.

Selected Projects

mycc: public C-subset compiler and reimplementation using Bison, Flex, and LLVM, with an optional ORC JIT path.

kscripT: Dynamic language implemented in C, with arrays, linear algebra, FFTs, and a WASM REPL.

CARVE: a collaborative RISC-V emulator and visualizer for the web, developed with Gregory Croisdale under Stephen Marz.

Blok: Built a C/C++ voxel game with world generation, chunk and mesh rendering, and input handling using OpenGL, GLFW, AssImp, FreeType, and PortAudio (2020).

Personal website and research blog: Technical writing, tutorials, and digital artwork, maintained since 2015.

Open Source Contributions

MPFR | Open Source Contributor | 2016–2017

Contributed arbitrary-precision implementations of scientific functions and assisted researchers through MPFR mailing lists.

Leadership

FIRST Robotics #3966 | Lead Programmer | 2015–2018

- Designed and implemented software for semi-autonomous robots competing in FIRST Robotics events.

Education

Bachelor's degree, Computer Science | University of Tennessee, Knoxville
2019–2023

- FIRST Robotics Alumni Scholarship