James **E. T. Smith, Ph.D.**

jamesETsmith / +1 508 596 7110 / james.smith9113@gmail.com /
james-smith-ph-d-8525792b / jamesetsmith.github.io/

Software engineer with experience in high performance computing, numerical linear algebra, and quantum chemistry, looking to work at the intersection of research and engineering teams.

Skills

Programming C/C++ (9+ years) Python (10+ years) CMake (8+ years) Rust (<1 year)
Parallelism ROCm OpenMP CUDA MPI
Tools git perf VTune rocprofv3 Slurm Cursor Crush

Professional Experience

Member of the Technical Staff
Jun. 2025 - Present
AMD — *Remote*

- Ported Deep Graph Library (DGL), a graph ML framework, from CUDA to ROCm for AMD datacenter GPUs.
- Accelerated ROCm ports of Quadrants and Genesis 3D physics libraries by 3x on AMD datacenter GPUs by improving the Quadrants JIT/LLVM pipeline to tag pointers with the correct AMDGPU address space, enabling faster global and shared-memory load/store paths.
- Tuned GEMM and kernel performance for Kimi K2.5 with ROCm's AITER library.
- Implemented the Step3.5 Flash model in ROCm's ATOM inference engine with 20-60% faster performance than vLLM across different server configurations.

C++ Python PyTorch ROCm HIP GPU Optimization Inference vLLM ATOM

Senior Research Engineer
Jan. 2024-Jun. 2025
Qognitive Inc. — *New York, NY*

- Engineering team lead after the departure of the CTO. Managed a team of 5 engineers across Machine Learning Engineering, Infrastructure, and Research.
- Managed the interview pipeline for two Applied Machine Learning Engineer positions, interviewed 25+ candidates, and hired 2 new engineers.
- Team leader for collaboration with IBM's Quantum Computing team. Worked alongside their research team to deploy our models on their quantum hardware and develop new model architectures to improve efficiency on their hardware.
- Lead developer for our open-source C++/Python project fast-pauli, which provides high-performance primitives for simulating Pauli-based quantum circuits.
- Worked between the machine learning research team and the infrastructure team to scale machine learning models from prototype to production
- Developed and implemented novel machine learning algorithms in Python, C++, and PyTorch for Qognitive's propriety SaaS product.
- Ported over existing models to PyTorch and optimized implementation for performance on GPUs yielding in a 10x speedup for several of our models.
- Contributed to the LLVM open source project by implementing std::ranges::iota (part of C++23 standard) in libc++, LLVM's version of the C++ standard library.

C++ C++23 Python PyTorch Machine Learning Quantum Computing Numerical Linear Algebra Linux

Member of the Technical Staff
Jul. 2022 - Jan. 2024
Lucata Corporation — *New York, NY*

- Developed a highly multithreaded version of the GraphBLAS library in C/C++, using C++17 and Cilk.
- Optimized the multithreaded performance of Lucata's GraphBLAS implementation by improving the utlization of the Lucata's proprietary hardware architecture.
- Collaborated with other teams regularly to address bugs and implement new features in the GraphBLAS and other graph analytics libraries.
- Overhauled the CMake infrastructure for the Lucata software ecosystem and set up continuous integration, testing, code coverage, and static linting for nearly all projects.
- Consolidated and improved the CMake build system for Lucata's custom LLVM 14 compiler.

C++ C++17 C multithreading Cilk CMake CI LLVM Numerical Linear Algebra Linux

Flatiron Research Fellow
Sep. 2020 - Jun. 2022
Center for Computation Quantum Physics, Flatiron Institute — *New York, NY*

- Implemented OpenMP parallelized stochastic compression methods for quantum chemistry in the open source C++ package FRI-CC which improved the asymptotic scaling of high-accuracy coupled-cluster methods
- Contributed features, bug fixes, and documentation as one of the primary maintainers for the open source electronic structure package, PySCF.
- Worked closely with the core team of PySCF developers improve the CMake build system and PyPI distribution after the release of PySCF v2.0.0.

- Organized workshops to help members of the Flatiron community better utilize high performance computing resources as part of the Sciware working group.

C++ C++20 Python multithreading OpenMP MPI CUDA CMake CI Linux

Graduate (Ph.D.) Researh Assistant *Aug. 2014 - Aug. 2020*
University of Colorado Boulder *Boulder, CO*

- Implemented a hybrid MPI-OpenMP parallelized version of the Heatbath Configuration Interaction (HCI) algorithm in the Sharma Group's C++ software Dice.
- Derived and implemented gradients of the HCI electronic Hamiltonian with respect to atomic positions enabling first-principles geometry optimization.
- Built decision tree and graph neural network models to predict etching reaction outcomes and trained these models with experimentally observed data.
- Frequently contributed to the open source PySCF quantum chemistry package, implementing new methods, features and handling bug reports.
- Wrote a new module for the PySCF package to interface with Dice enabling the investigation of previously intractable systems.
- Organized and led a workshop on software best practices for graduate students and post doctoral researchers with staff from the Molecular Sciences Software Institute (MOLSSI).

Python C++ C++11 multithreading OpenMP MPI CMake machine learning scikit-learn Linux

Education

Ph.D. Chemical Physics *Aug. 2014 - Aug. 2020*
University of Colorado Boulder *Boulder, CO, US*

Bachelors of Science in Chemistry, Minor in Math *Aug. 2010 - May 2014*
Davidson College *Davidon, NC, US*

Volunteer

Instructor *May 2021 - Present*
Software Carpentry *New York, NY*

- Taught regularly about software best practices in scientific computing to learners with a broad programming background. Taught lessons on shell, Git, Python, and data visualization in Python.

Certifications

May 2021 **Software Carpentry Instructor Certificate**, Software Carpentry
Apr. 2021 **NVIDIA DLI Certificate - Accelerating CUDA C++ Applications with Multiple GPUs**, NVIDIA
Apr. 2021 **NVIDIA DLI Certificate - Fundamentals of Accelerated Computing with CUDA C/C++**, NVIDIA