[redacted]

Pate Motter

[redacted]

Education

- **University of Colorado, Boulder** — Boulder, CO
  Doctor of Philosophy in Computer Science — *Aug. 2017*
  - **Dissertation Title**: Hardware Awareness for the Selection of Optimal Iterative Linear Solvers

- **University of Colorado, Boulder** — Boulder, CO
  Master of Science in Computer Science — *May 2013*

- **University of Arkansas, Fayetteville** — Fayetteville, AR
  Bachelor of Science in Computer Science, minor in Mathematics — *May 2011*

Work and Research Experience

- **Google** — Seattle, WA
  Software Engineer, TPU Inference — *May 2023 – Present*
  - Architected and implemented Paged Attention for MaxText from design to merge, resolving complex concurrency bugs and achieving a 47% throughput increase in microbenchmarks.
  - Developed custom Pallas kernels for 2D block-wise and sub-channel quantization, enabling advanced precision support within the TPU-Inference framework on TPU.
  - Built model accuracy benchmarking framework for tracking long-term improvements and regressions of TPU-Inference models.

- **Amazon (Alexa AI)** — Seattle, WA
  Machine Learning Engineer, Web-based Question-Answering — *Oct 2021 – May 2023*
  - Reduced overall latency of our model by 6x compared to the model's baseline on GPUs.
  - Responsible for performance analysis and optimization of a transformer-based research deep learning model.
  - Created codebase that compiles our model to ONNX, TensorRT, and Inferentia formats and runs benchmarking.

- **Amazon Web Services (AWS HPC)** — Seattle, WA
  Research Engineer, Application Performance Team — *March 2020 – May 2021*
  - Created automated performance regression testing system for AWS HPC infrastructure.
  - Benchmarked prototype EC2 instances to identify optimal hardware configurations for future AWS HPC EC2 instances.

- **Amazon Alexa** — Seattle, WA
  Research Engineer, Automatic Speech Recognition — *March 2018 – March 2020*
  - Created a new team to maintain an in-house deep learning framework and address organizational needs.
  - Optimized distributed programming workflows using C++, Python, CUDA, TensorFlow, and MxNet.

- **Amazon Web Services** — Seattle, WA
  Software Development Engineer, Mobile Hub — *Aug 2017 – March 2018*
  - Built Java-based backend services and data collection pipelines for customer usage analytics.

- **University of Colorado, Lighthouse Project** — Boulder, CO
  Doctoral Research / Research Assistant — *Aug 2014 – Aug 2017*
  - Utilized runtime performance data from supercomputers to train ML algorithms that predict optimal iterative linear solvers (C++, Trilinos, Python).

- **Lawrence Livermore National Laboratory** — Livermore, CA
  Computation Intern — *Summer 2014, 2015*

- ○ Optimized BLAST hydrodynamics code for future architectures and developed benchmarking suites for high-performance linear algebra libraries using HPCToolkit.

- **TerraSpark Geosciences** Boulder, CO
  Software Developer / Researcher *Aug 2011 – Jan 2014*
  - ○ Implemented GPU-based seismic interpretation solutions using OpenCL, reducing processing time from hours to seconds compared to original CPU code.

Patents

- **US-12093669-B1**: Massively parallel compilation of application code (Issued 2024-09-17)

Selected Publications & Awards

- **Publication**: E. Jessup, P. Motter, et al., ”Performance-Based Numerical Solver Selection in the Lighthouse Framework,” SIAM Journal on Scientific Computing, 2016.
- **Publication**: K. H. Koh... P. Motter, ”Will it stick? exploring the sustainability of computational thinking education,” SIGCSE 2013.
- **Award**: Amazon ”Puzzle Piece” for submitting a patent to the US patent office (2021)
- **Award**: XSEDE Allocation ($200k+ SUs) for furthering my thesis work - #CCR160019 (2016)
- **Award**: Nvidia Research Grant for GPU-based machine learning and numerical linear algebra (2015)

Skills

- **Languages**: Python, C++
- **Libraries & Frameworks**: JAX, Pallas, CUDA, MPI, OpenMP, OpenCL
- **Tools**: XProf, DDT, ARM MAP, Intel Vtune, HPC Toolkit