Daniel Zou

(703) 939-0901 | dzou@nvidia.com | dlzou | dlzou | dlzou.github.io

Education

University of California, Berkeley Berkeley, CA
B.Sc. Electrical Engineering and Computer Sciences, with High Honors; 3.95/4.00 GPA *Aug 2019 – May 2023*

- Selected coursework: Algorithms, Compilers, Parallel Computing, Computer Architecture, Operating Systems, Databases, Machine Learning, Deep Learning, Signal Processing

Stanford Center for Professional Development Stanford, CA
Sponsored Enrollment in Graduate Level Coursework *Jan 2024 – Present*

- Coursework: Program Analysis, Deep Learning for NLP, Robot Autonomy I & II

Work Experience

NVIDIA Corporation Santa Clara, CA
Systems Software Engineer, AI Infrastructure *Jun 2023 – Present*

- Worked on pipelines that generate ground truth labels for multi-modal sensor signals for AV model training.
- Led initiative to quantify GPU cost and inefficiencies in compute-intensive pipelines; built analytics infrastructure and introduced the Nsight Systems profiler tool.
- Through profiler-guided optimization, mitigated memory and IO bottlenecks in critical egomotion pipeline to reduce consumed GPU hours by about 20%, or 2,500 DGX node hours per month.
- Developed QA pipeline for lidar-camera calibration that intelligently selects highly representative data samples to reduce human review time.
- Worked on backend services for managing the curation and assembly of datasets with millions of assets.
- Responsible for redesigning several service components to migrate to a new data lake backend.

NVIDIA Corporation Santa Clara, CA
Autonomous Vehicles Software Infrastructure Intern *May 2022 – Aug 2022*

- Designed a application for automated management and metrics collection in on-call channels, which are used by all autonomous vehicle infrastructure teams.
- Implemented a full stack event-driven service in Go that concurrently tracks status and statistics for hundreds of live support threads. Statistics were persisted in a relational database and were queryable through a chat interface.
- Worked with multiple stakeholders to revise features and improve usability of the application during development and beta testing. Deployed for use by over 2,000 engineers and business partners.

Research Experience

Berkeley Sky Computing Lab Berkeley, CA
Undergraduate Research Assistant *Sep 2022 – Dec 2022*

- Contributed to Alpa, a framework based on JAX and Ray that automatically parallelizes the training and serving of neural networks with hundreds of billions of parameters.
- Integrated OPT-IML models into the Alpa framework and did performance benchmarking.
- Developed a proof-of-concept runtime that follows an optimal policy to swap tensors between GPU and main memories, enabling deployment on hardware with less GPU resources.
- Advised by Lianmin Zheng and Professor Joseph E. Gonzalez.

Berkeley RISE Lab Berkeley, CA
Undergraduate Research Assistant *Mar 2021 – Aug 2022*

- Contributed to NumS, a high performance distributed numerical library for Python with a NumPy-like interface, a Ray backend, and a novel simulation-based scheduler.
- Introduced support for sparse tensors, then studied how to use operator fusion and cost estimation to dynamically schedule sparse operations on clusters in a memory-efficient manner.
- Implemented the distributed quickselect algorithm for block-partitioned arrays, then verified strong and weak scaling through benchmarks of up to 2^{33} double precision floats on an AWS EC2 cluster with up to 256 worker processes.
- Advised by Melih Elibol and Professor Ion Stoica.

Selected Projects

Computron: Serving Distributed Deep Learning Models with Model Parallel Swapping (ArXiv)

- Built a deep learning inference system that serves multiple distributed models on a shared GPU cluster.
- Designed a distributed model swapping mechanism that takes advantage of the aggregate PCIe bandwidth of a multi-GPU cluster to reduce latency, making it feasible to oversubscribe GPU memory with large models.
- Evaluated the system by serving large transformer-based models such as OPT and ViT on a single node cluster with up to four A100 GPUs.
- Independent research project for CS 267: Applications of Parallel Computers, advised by Professors James Demmel and Hao Zhang.

CycleGAN-JAX (GitHub)

- Reimplemented the CycleGAN image generation architecture using the JAX framework.
- Trained and evaluated image-to-image style transfer on several datasets such as horse2zebra; achieved outputs that are qualitatively on par with the reference implementation.
- Independent final project for CS 182: Deep Learning.

Volunteer Experience

Computer Science Undergraduate Association Berkeley, CA
VP of Technology & Root Staff *Sep 2020 – Dec 2022*

- Administered a GPU cluster, a web hosting server, and other free computing services used by 400+ members.
- Maintained a full stack web application written in Django.
- Configured a multi-instance Postfix mail server.
- Served as the leader of CSUA Root Staff for the Fall 2021 semester upon election to VP of Technology. Acted as liaison with campus IT staff to resolve issues regarding networking services.

Computer Science Mentors Berkeley, CA
Junior Mentor *Jan 2021 – May 2021*

- Taught CS 61C: Great Ideas in Computer Architecture, which covers C, RISC-V assembly, CPU logic design, memory hierarchy, and parallel programming.
- Mentored a group of students every week by giving short lectures and practice problem walk-throughs.

Skills

Languages: Python, C/C++, CUDA, Go, Rust, Java, OCaml, Lua, SQL, RISC-V, x86

Libraries: NumPy, PyTorch, JAX, Ray, Spark, ROS, MPI

Technologies: Ubuntu Linux, Docker, Bazel, CMake, NVIDIA Nsight

Other: native English and Mandarin Chinese, technical and blog writing