Brian Hsu

43757 Excelso Drive, Fremont, CA 94539

(510) 493-8123 | brian@brianhsu.me | brianhsu.me | brianhsu98 | brianhsu98

Work Experience

OpenAI
San Francisco, California

Member of Technical Staff — *April 2024 - present*

- Building infrastructure to run GPUs at scale.

Databricks
Mountain View, California

Senior Software Engineer — *Sept 2022 - April 2024*

- Engineer on the Compute Infra team, building a highly scalable, efficient, and easy-to-use Kubernetes platform for all internal teams at Databricks.
- **Owned compute budget (tens of millions) from end-to-end**, ultimately owning cost management for all services running in the control plane (supporting most of the companys' revenue). **Set overall roadmap** for cost efficiency efforts, drove quarterly budget forecasting process, built systems to attribute cost and drive budget adherence from a service level, and delivered on cost savings totaling millions of dollars each year.
- Drove horizontal autoscaling company-wide **from 0% adoption to cover the majority of critical, eligible services**. Worked directly with service teams as an autoscaling expert to configure their services optimally, saving on cost while also protecting services from overload.
- Designed and built our **next generation, highly performant autoscaling system**, offering low-latency (sub-second) scale-ups for customers' inference workloads.

Meta
Menlo Park, California

Production Engineer — *Feb 2020 - Sept 2022*

- Senior engineer on the Resource Allowance System team, **designing and implementing capacity allocation workflows** for **thousands of customers** across **millions of machines**, providing the foundation for Meta's internal cloud.
- Helped develop **IaaS Experimentation**, a system for users to acquire hardware, run containers, and apply custom automation for the purpose of testing different workloads on different hardware. Onboarded and supported customers, designed and implemented features, and came up with new projects for other team members.
- Designed and implemented systems to **automatically distribute and reclaim servers** from internal customers, providing them with necessary fault-tolerance buffer, along with improving fleet spread and hardware scheduling.
- **Mentored junior members** of my team and came up with a large variety of projects to aid in their engineering development.

Algorithms for Computing and Education (ACE) Lab, UC Berkeley
Berkeley, California

Research Assistant — *May 2018 - Dec 2019*

- Worked with PhD student Nate Weinman, advised by Professor Armando Fox, to **research and develop novel computer science practice problems** to make computer science more accessible and easier-to-learn for beginning and intermediate students.
- Collaboratively designed and implemented an **interactive web application** to solve Parsons Problems, enabling a 80+ student research study, along with a **parallelized autograding** system.

LiveRamp
San Francisco, California

Software Engineering Intern, Data Management Backend — *May 2019 - Aug 2019*

- Built big data systems, helping **add to, segment, and process petabytes of customer data** to enable data-driven marketing.
- Collaborated across teams, **implementing new endpoints** to enable easier access to my team's systems. Developed and owned a new backend service and big data pipeline.

Education

University of California, Berkeley
Berkeley, CA

B.A. in Computer Science, Minor in English. Major GPA 3.80, Cumulative GPA 3.65 — *Aug. 2016 - Dec. 2019*

Skills

Programming: Rust, C++, Python, Java, JavaScript, Go, C, SQL

Technologies: IaaS, Kubernetes, Autoscaling, Systems Performance, Distributed Systems, Cloud Cost, Containerization