Hariharan Sezhiyan

US Citizen | hs86@rice.edu | 510-456-8704 | https://hsezhiyan.github.io/

Research Experience

Rice University

- Currently a 2nd year PhD student working with Prof. Eugene Ng and Prof. Ang Chen (at the University of Michigan)
- My research consists of improving machine learning systems across all layers of the systems stack, from the host layer to the networking layer.
- Working on a project to improve the memory efficiency of deep learning recommendation models (DLRMs) by leveraging unique data access patterns of embedding vectors in huge embedding tables. The end goal is to deliver a system that is more adaptable to the cost and performance needs of the customer.
- Working on a project to enable network topology and multitenant aware deep learning training on the public cloud. For this project, I am working on (1) enabling distributed deep learning frameworks like Horovod and DeepSpeed to be aware of the network topology of the compute cluster they are running on, and (2) enabling network performance improvements by exploiting the fact that some network flows in AI clusters can be deprioritized without any performance penalties to that job. The end goal is to deliver a new paradigm in ML computing on the cloud which is more performant and gives more confidence to model developers.

UC Davis

- Worked at the UC Davis DECAL lab with Prof. Premkumar Devanbu on software engineering research
- Worked on an empirical software engineering project to analyze code to comment datasets. Publication: Code to Comment "Translation": Data, Metrics, Baselining & Evaluation, Accepted at the 36 th IEEE/ACM International Conference on Automated Software Engineering **(shared first author)**
- Worked on transformer models to generate comments from code

Work Experience

Amazon Web Service (AWS) | Software Engineer II | Sunnyvale, CA — February 2022 - December 2022

- Worked on the EC2 dataplane team
- **Developed** a new ingestion service to aggregate telemetry data from EC2 across all hosts across all AWS regions. This ingestion service feeds to other internal tools and dashboards. Daily ingestion volume exceeds 10 terabytes of data.
- **Developed** a new dashboard service called Interface Analyzer which allows internal users to extract various networking properties of EC2 hosts across the entire AWS fleet. This tool has been instrumental in mitigating outages and addressing external customers' concerns, with average time to resolution going from 2 weeks to 2 days.

Zillow Group | Machine Learning Engineer II | San Francisco, CA — July 2020 - February 2022

- **Developed and owned** a **distributed** integration testing framework for the AI Platform Workflow SDK, which allows applied scientists to easily run pipelines on **Kubeflow Pipelines**. Runs a set of tests in parallel on the AI Platform's internal **EKS** cluster. Instrumental for the successful delivery of the Workflow SDK production release. Also **developed parts** of the Workflow SDK.
- **Developed and owned** a **Spark** SDK as part of the Workflow SDK, allowing data engineers to easily run Spark on Kubernetes. Used by 10 internal AI/data engineering teams.
- **Owned** the entire integration testing space for the AI Platform. Responsible for integration tests for: Workflow SDK, Spark on **EKS**, **Distributed PyTorch Training**, **Kubeflow Serving**.
- **Mentored** interns as part of the company intern-engineer matching program.

Innovation Minds | Software Development Engineer | San Jose, CA — January 2018 – June 2020

- **Created** scalable and extensible web infrastructure for the Innovation Minds product, including collaborative tools and event/idea management services. Technologies used: Node.js, Java, AWS Redshift, and Python.
- **Deployed** all services to production environments and maintained AWS EKS clusters for scaling. Used by 100 clients.

Education

- Rice University, 2nd year PhD Student, 4.0 GPA **(ongoing, expected to graduate around 2027)**
- UC Davis, BS Computer Science and Engineering, Mathematics Minor. 3.8 GPA, Magna Cum Laude

Technical Skills

- Experienced: Distributed Systems, Python, Go, Docker, Kubernetes, Kubeflow, Metaflow, CICD, C++, C
- Exposure: Spark, Hadoop, Kafka, Airflow, TensorFlow/PyTorch, Data Engineering, NLP for source code