Vikranth Srivatsa

vsrivatsa@berkeley.edu ◇ github.com/vikranth22446 ◇ viks.me

EDUCATION

University of California San Diego - Ph.D. Oct 2023 - May 2027
Computer Science: Systems and Machine Learning Advised by Yiying Zhang in WukLab

University of California Berkeley - Bachelor & Masters of Science Aug 2018 - May 2023
Electrical Engineering and Computer Science Advised by Joseph Gonzalez in Rise/Sky Lab

Skills
Languages: Python, Rust, Golang, C++, Java, C, Scala, Verilog, Risc-V/x86, CUDA
Other: Pytorch, Tensorflow, Natural Language Processing, Keras, OpenCV
Interest: Efficient LLM, LLM Inference, Distributed Systems, MLSys

RESEARCH EXPERIENCE

→ **DAS: Distribution-Aware Speculative Decoding for RL Post-Training @ Together AI**
MLSys 2026 Jun 2025 - Dec 2025
- Research internship at Together AI: designed DAS, a speculative decoding framework that accelerates RL rollout generation for LLMs (co-first author, with Tri Dao, Ce Zhang, Percy Liang)
- Proposed a length-aware speculation policy that prioritizes aggressive decoding on long-tail trajectories to reduce rollout makespan
- Achieved up to 50% reduction in rollout latency on agentic reasoning workloads with no change to accuracy

→ **TClone: Forkable Personal Workspaces for Computer-Use Agents** Jan 2025 - Sep 2025
- Designed an efficient method to make copies of containers for RL/computer use
- Built low-latency workspace versioning using Linux kernel modifications, CRIU extensions, copy-on-write memory sharing, and filesystem versioning
- Evaluated TClone across 600+ agent tasks, achieving up to 1.9× lower end-to-end task latency compared to KVM-based baselines

→ **Adaptive Parallelism Multi SLO Serving @ WukLab** Sept 2024 - April 2025
- Building a system that dynamically changes level of tensor parallelism to meet strict SLOs
- Wrote custom CUDA Kernels to optimize performance for specific SLOs
- Discovered the system problem and led a team of 5 engineers to work on the project.

→ **Efficient Distributed Prompt Scheduling for LLMs @ WukLab** *ICLR 2025* - Aug 2024
- Built a system for efficient LLM load balancing resulting in 1.5X to 14.5X on average latency and 2X to 10X on p99 latency over SOTA
- Utilizes prefill decode balancing, traditional load balancing/caching, and fairness techniques to achieve optimal performance.
- Investigated and discovered the problem of distributed prefix caching

→ **Efficient Augmented Language Models @ WukLab** *ICML 2024* Sept 2023 - July 2024
- Built a inference system that improves latency/throughput of language models that call APIs
- Built load testing tool profiling different augmented language models vLLM, OpenAi, ToolFormer

→ **NASA AMES and SkyLab** June 2021 - July 2023
- Created a serverless scheduling framework across client, 5G networks, & cloud (AWS/GCP/Azure/K8s)
- Wrote heuristic and linear programming to save 70% cost reduction and 30% performance increase
- Profiled and optimized throughput, implemented write ahead logging and fault tolerance

→ **UC Berkeley RISELab** *ML Researcher - NeurIPS 2021 Workshop* Feb 2021 - Dec 2023

- Investigating machine learning explainability with large models and distribution shift
- Designed and ran experiments to test model size, subgroup robustness and counts, and pretraining
- Automatically constructed spurious correlation large scale image datasets with Imagenet

WORK EXPERIENCE

→ **MLSys Researcher UCSD** June 2023 - Present
- Designed and built an efficient distributed LLM scheduling system on top of vLLM and SGLang
- Led problem discovery to optimize LLM scheduling, while managing a team of 5+ engineers to implement solutions.

→ **New Relic Pixie Labs** *Systems Software Engineering Intern* June 2022 - May 2022
- Designed a new system that automatically parses system level protocols used in K8s in C++
- Architected RabbitMQ and Kafka monitoring and performance analysis tooling in Kubernetes
- Worked on kernel level memory profiling and flame graph creation via kernel monitoring protocol eBPF

→ **Amazon Alexa** *Software Development Engineer Intern* May 2020 - August 2020
- Designed and built dev tool for experiment tracking, managing A/B model testing, and dashboards
- Built parameter tracking to process models used in ranking content in the Alexa Mobile App

→ **Amazon AWS** *Software Development Engineer Intern* May 2019 - August 2019
- Implemented core open source features on AWS SAM-CLI used in 13% of production workflows
- Designed cloudformation and cloudformation intrinsic templating resolution
- Built a Serverless Testing Framework to benchmark and test AWS Serverless infrastructure

→ **Circles - Startup** *Chief Technology Officer* September 2018 - May 2019
- Lead a team of 10 engineers to build out a curated news platform
- Created an efficient news recommendation system using feed ranking algorithms and collaborative filtering

PUBLICATIONS

→ **Nitsum: Serving Tiered LLM Requests with Adaptive Tensor Parallelism**
arXiv 2605.05467 May 2026
Vikranth Srivatsa, Zijian He, Pu Guo, Dongming Li, Yiying Zhang

→ **TClone: Low-Latency Forking of Live GUI Environments for Computer-Use Agents**
arXiv 2605.17320 May 2026
Yutong Huang, **Vikranth Srivatsa**, Alex Asch, Hansin Tushar Patwa, Yiying Zhang

→ **Beat the long tail: Distribution aware spec decoding for rl training**
MLSys 2026 May 2026
Zelei Shao*, **Vikranth Srivatsa***, Sanjana Srivastava, Qingyang Wu, Alpay Ariyak, Xiaoxia Wu, Ameen Patel, Jue Wang, Percy Liang, Tri Dao, Ce Zhang, Yiying Zhang, Ben Athiwaratkun, Chenfeng Xu, Junxiong Wang

→ **Cognify: Supercharging Gen-AI Workflows With Hierarchical Autotuning**
KDD 2026 Feb 2025
Zijian He*, Reyna Abhyankar*, **Vikranth Srivatsa**, Yiying Zhang

→ **Preble: Efficient Distributed Prompt Scheduling for LLM Serving**
ICLR 2025 May 2024
Vikranth Srivatsa*, Zijian He*, Reyna Abhyankar, Dongming Li, Yiying Zhang

→ **InferCept: Efficient Intercept Support for Augmented Large Language Model Inference**
ICML 2024 July 2024

Zijian He*, Reyna Abhyankar*, **Vikranth Srivatsa**, Hao Zhang, Yiying Zhang

→ **Cloudless Serverless across multi cloud and Mixclaves: secure private messaging**
UC Berkeley EECS Thesis 2023
Vikranth Srivatsa, Alan Pham, Moustafa AbdelBaky, Gabe Fierro, John Kolb, Joseph Gonzalez David Culler
Vikranth Srivatsa, Chris Douglas, Mark Theis, John Kubiatowicz, Arjit Panda, Scott Shenker.

→ **Effect of Model Size on Worst-group Generalization**
Neurips 2021 Workshop 2021
Vikranth Srivatsa*, Alan Pham*, Eunice Chan*, Dhruba Ghosh, Yaoqing Yang, Yaodong Yu, Ruiqi Zhong, Joseph E. Gonzalez, Jacob Steinhardt

→ **Passive Listening Smart Speakers - Berkeley Security Lab**
Published at Center of Long Term Cybersecurity 2022
Vikranth Srivatsa, Bhuvan Basireddy, Hrithik Datta, Prakash Srivastava, and Varun Jadia, Nathan Malkin, Serge Egelman

PROJECTS

→ **CPU Scheduling Overhead Analysis in vLLM and SGLang - @Wuklab** August 2024
- Investigated cpu scheduling overhead in vLLM, helping vLLM re-design their system for 30% better performance.
- Published as a popular Blog Post `https://mlsys.wuklab.io/posts/scheduling_overhead`

→ **Passive Listening Smart Speakers - Berkeley Blues Security Lab** Feb 2020 - Feb 2021
- Created a system to study privacy implications of always listening smart assistants via ML
- Built machine learning NLP models to classify different intentions for passive listening and speech
- Published at UC Berkeley Center for Longterm Cybersecurity