Ahan Gupta

Email : ahangupta.96@gmail.com
Mobile : +1-415-966-5501
Linkedin
Github: https://github.com/spikerheado1234

Education

- **University of Illinois Urbana-Champaign** — Champaign, IL
  PhD in Computer Science — *Aug 2022 - Present*
- **National University of Singapore** — Singapore
  Bachelor of Computing in Computer Science — *Aug 2017 - Dec 2021*

Research Statement

I am broadly interested in researching high-performance Compiler & System level abstractions to accelerate deep-learning applications. My work melds both theory and practice, providing high-performance abstractions and systems that have strong theoretical guarantees.

Experience

- **Google DeepMind** — Mountain View, CA
  Student Researcher — *May 2024 - November 2024*
  - Investigated trainable KV-cache compression strategies at scale.
- **Citadel** — Hong Kong
  Software Engineering Intern — *May 2021 - Aug 2021*
  - Designed an authentication library to enable developers to integrate authentication logic with different services
  - Contributed to a tool that monitors AWS usage of different desks
  - Designed and built a monitoring tool that enables traders to track internal services' uptime and accuracy
- **Google** — Singapore
  Software Engineering Intern — *May 2020 - Aug 2020*
  - Designed Asynchronous Web APIs via OpenAPI for authorisation microservice in MojaLoop network
  - Designed database Schemas & built infrastructural groundwork to enable integration with said databases
  - Implemented APIs that enable secure FIDO signature validation in HapiJS and TypeScript
  - Merged all code into production

Publications

- **Ahan Gupta**, Yueming Yuan, Devansh Jain, Yuhao Ge, David Aponte, Yanqi Zhou, Charith Mendis. SPLAT: Optimized GPU code generation framework for SParse reguLar ATtention. OOPSLA 2025.
- Hoa La*, **Ahan Gupta***, Alex Morehead, Jianlin Cheng, Minjia Zhang. MegaFold: System-Level Optimizations for Accelerating Protein Structure Prediction Models. In submission 2025.
- Yueming Yuan, **Ahan Gupta**, Jianping Li, Sajal Dash, Feiyi Wang, Minjia Zhang. X-MoE: Enabling Scalable Training for Emerging Mixture-of-Experts Architectures on HPC Platforms. In submission 2025.
- Muyan Hu, **Ahan Gupta**, Jiachen Yuan, Vima Gupta, Xin Xu, Janardhan Kulkarn, Ofer Dekel, Vikram Adve, Charith Mendis. VTC: DNN Compilation with Virtual Tensors for Data Movement Elimination. In submission 2025.
- **Ahan Gupta**, Hao Guo, Yueming Yuan, Yanqi Zhou, Charith Mendis. FLuRKA: Fast fused Low-Rank & Kernel Attention. In Submission 2025. Preprint link: https://arxiv.org/abs/2306.15799

* Denotes Equal Contribution

Service

- **ACM TACO Reviewer**: 2025
- **ISCA AEC**: 2024

Skills Summary

- **Languages**: Java, C++, Python, C, SQL, Javascript, Scala, Cuda
- **Tools**: Docker, Pytorch, Tensorflow, JAX, LLVM