

Ahan Gupta

Linkedin

Github: <https://github.com/spikerheado1234>

Email : ahangupta.96@gmail.com

Mobile : +1-415-966-5501

EDUCATION

- | | |
|--|----------------------------|
| University of Illinois Urbana-Champaign | Champaign, IL |
| • <i>PhD in Computer Science - Advisor: Minjia Zhang</i> | <i>Aug 2022 - Present</i> |
| National University of Singapore | Singapore |
| • <i>Bachelor of Computing in Computer Science</i> | <i>Aug 2017 - Dec 2021</i> |

EXPERIENCE

- | | |
|---|---------------------------------|
| Google DeepMind | Mountain View, CA |
| • <i>Student Researcher</i> | <i>May 2024 - November 2024</i> |
| ◦ Investigated novel low-rank compression techniques to reduce KV-cache sizes for 10B+ parameter LLMs. | |
| Citadel | Hong Kong |
| • <i>Software Engineering Intern</i> | <i>May 2021 - Aug 2021</i> |
| ◦ 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 |
| • <i>Software Engineering Intern</i> | <i>May 2020 - Aug 2020</i> |
| ◦ 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

- Muyan Hu, Ahan Gupta, Jiachen Yuan, Vima Gupta, Taeksang Kim, Xin Xu, Janardhan Kulkarni, Ofer Dekel, Vikram Adve, Charith Mendis. VTC: DNN Compilation with Virtual Tensors for Data Movement Elimination. OSDI 2026.
- 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. SC 2025. **Best Student Paper**.
- 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.
- Ahan Gupta*, Zhihao Wang*, Neel Dani, Masahiro Tanaka, Olatunji Ruwase, Minjia Zhang. AutoSP: Unlocking Long-Context LLM Training Via Compiler-Based Sequence Parallelism. In submission 2025.
- Hoa La*, Ahan Gupta*, Alex Morehead, Jianlin Cheng, Minjia Zhang. MegaFold: System-Level Optimizations for Accelerating Protein Structure Prediction Models. 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