Abhishek Vijaya Kumar

Ph.D. candidate, *Cornell University* abhishek@cs.cornell.edu
www.cs.cornell.edu/ abhishek/

RESEARCH INTERESTS

LLM Reinforcement learning Infrastructure, ML systems and scheduling for LLM serving. Communication optimizations for distributed machine learning with a focus on GPU-interconnects. Recently working on reducing overheads of calling external tools in RL post training and inference, including post-training LLMs with **10–30B parameters**. Research has lead to publications at **ICML, NSDI, ASPLOS, FAST, OFC and Hotnets**.

EDUCATION

Ph.D., Computer Science, **Cornell University**, **2021-August 2026** (**expected**), Advisor: Rachee Singh

B.E., PES Institute of Technology (PESIT), 2014-2018, CGPA 9.31, *Department Rank **6** in a batch of **81***

WORK EXPERIENCE

- **Microsoft Research Redmond** *Summer 2022, Summer 2024 — Redmond, WA*
  Research Intern, Networking Research Group
  - Worked on GPU and external memory interconnect optimization for LLM inference.
  - Developed tools to diagnose issues in large-scale kubernetes systems.

- **Microsoft Research India** *Jul 2018-Jan 2021 — Bengaluru, India*
  Research Fellow (Full-time) (2018–2021), Research Intern (2016, 2017), Systems and Networks Group
  - Designed storage caching (Quiver) for distributed ML in cloud.
  - Built large-scale data pipelines and mobile sharing networks.

- **Intuit India** *Jan'18 – Jun'18 — Bengaluru, India*
  Software Development Intern
  - Worked on backend and frontend of QuickBooks Online to complement UAE's tax scheme.

AWARDS AND SCHOLARSHIPS

- ACE-SRC best poster award for LLM inference work (Aqua).
- Bowers-LinkedIn Fellowship for the academic year of 2024-2025.
- Outstanding TA award for Operating Systems (Fall 2021, Fall 2022), Cornell University.
- CNR Merit Scholarship from PESIT every year (awarded to top 20% students).
- MRD Scholarship from PESIT during two semesters (top 5 ranks).
- Won 2nd place in Microsoft internal worldwide hackathon.

SELECTED RESEARCH

TVCACHE: Reducing tool calling overheads in RL fine-tuning for LLMs Cornell, Present
Fine-tuning LLMs to call tools requires executing the tools in an execution environment. TVCACHE reduces environment initialization overheads and tool calling overheads using a stateful and accurate cache for tool calls and environments, evaluated while RL post-training LLMs with **10–30B parameters**. [ICML 2026]

Aqua: Making LLM inference more responsive during traffic surge Cornell, April 2025
Makes LLM inference more responsive during unexpected surge of users by harvesting unused memory for fast preemption on interconnected GPUs over NvLinks. Aqua reduces the time-to-first-token (TTFT) by several seconds. [ASPLOS 2025, code].

Reconfigurable Torus Fabrics for Multi-tenant ML Cornell, August 2025
Prototyped a server-scale programmable direct-connect photonic fabric to interconnect accelerators within servers. Reconfigurable Torus Fabrics can improve the bandwidth of tenant compute allocations by up to 66%, reduce compute fragmentation by up to 70%, and minimize the blast radius of chip failures. [Hotnets 2024, ASPLOS 2026]

Arna: Accelerating GPU Access to External Memory for Large Models Microsoft Research, Summer 2024
Many LLM workloads resort to offloading data to slower memory like DRAM to manage memory capacity bottlenecks. Designed a multipath routing algorithm within the scale-up domain at Microsoft Research that linearly scales GPU-to-DRAM bandwidth with GPU hops.

PUBLICATIONS

TVCACHE: A Stateful Tool-Value Cache for Post-Training LLM Agents
Abhishek Vijaya Kumar*, *Bhaskar Kataria**, *Byungsoo Oh*, *Emaad Manzoor*, *Rachee Singh*
International Conference on Machine Learning (ICML) 2026. *Co-first authors.

Reconfigurable Torus Fabrics for Multi-tenant ML
Abhishek Vijaya Kumar, *Eric Ding*, *Arjun Devraj*, *Darius Bunandar*, *Rachee Singh*
Architectural Support for Programming Languages and Operating Systems (ASPLOS) 2026

Aqua: Network-Accelerated Memory Offloading for LLMs in Scale-Up GPU Domains
Abhishek Vijaya Kumar, *Gianni Antichi*, *Rachee Singh*
Architectural Support for Programming Languages and Operating Systems (ASPLOS) 2025

A case for server-scale photonic connectivity
Abhishek Vijaya Kumar, *Arjun Devraj*, *Darius Bunandar*, *Rachee Singh*
HotNets 2024

Chip-to-chip photonic connectivity in multi-accelerator servers for ML
Abhishek Vijaya Kumar, *Arjun Devraj*, *Darius Bunandar*, *Rachee Singh*
Optical Fiber Communication Conference (OFC) 2025

Chisel: An optical slice of the wide-area network
Abhishek Vijaya Kumar, *Bill Owens*, *Nikolaj Bjoner*, *Binbin Guan*, *Yawei Yin*, *Victor Bahl*, *Rachee Singh*
USENIX Symposium on Networked Systems Design and Implementation (NSDI) 2024

Quiver: Informed storage caching for deep learning training jobs
Abhishek Vijaya Kumar, *Muthian Sivathanu*
USENIX Conference on File and Storage Technologies (FAST) 2020

PCCL: Photonic circuit-switched collective communication for distributed ML
Abhishek Vijaya Kumar, *Arjun Devraj, Rachee Singh*
Under submission, available on arXiv 2025

Efficient AllReduce with Stragglers
Arjun Devraj, Eric Ding, ***Abhishek Vijaya Kumar****, Robert Kleinberg, Rachee Singh*
Under submission, available on arXiv 2025

TECHNICAL SKILLS

Frameworks: vLLM, Sglang, PyTorch, VeRL, Tinker
Languages: Python, C++, C, CUDA
Skills: Internals of ML training and inference frameworks, post-training systems for LLMs, collective communication algorithms.

TEACHING EXPERIENCE

- TA for CS 4450 (Introduction to Computer Networks), Spring 2023, Spring 2021.
- TA for CS 4410 (Operating Systems), Fall 2021, Fall 2022.

REFERENCES

- Rachee Singh, PhD advisor: rachee@cs.cornell.edu