Aashaka Shah

Senior Researcher
Microsoft Research, Redmond

aashakads@gmail.com
https://aashaka.github.io

Education

University of Texas at Austin **2018-2023**
Ph.D. in Computer Science
Advisor: Prof. Vijay Chidambaram

Indian Institute of Technology Roorkee **2014-2018**
B.Tech in Computer Science and Engineering

Areas of Interest

Building high-performance resource-efficient systems for LLM inference and RL post-training, particularly for long-horizon agentic tasks. Network, memory, KV cache, scheduling, and resource allocation optimizations for large-scale deep learning. Communication libraries for distributed ML

Tools, Libraries, and Languages: PyTorch, vLLM, FSDP, Megatron, verl, rllm, Nvidia Nsight, NCCL, MSCCL++, Python, Ray, C++, CUDA, Gurobi, Z3

Work Experience

Microsoft Research, Redmond
Senior Researcher Aug 2023-present

- Currently leading system optimization effort for RL post-training for LLMs at MSR Redmond
- Focus on workload-specific optimizations for agentic post-training, resource allocation, and optimization of different collective communication patterns
- Previously worked on scheduling and KV cache management for efficient LLM inference

Meta AI Systems HW/SW Co-design - SWE Intern May - Aug 2022
Adding vector semantics for ReduceScatter and Allgather collective to PyTorch and NCCL

Microsoft Research, Redmond - Research Intern May - Aug 2021
Synthesizing collective communication algorithms for a distributed GPU network

Microsoft Research, India - Research Intern May - Aug 2019
Succinct verifiable computation and zero-knowledge proofs

Adobe Research, India - Research Intern May - July 2017
User-side noisy-neighbour handling for Apache Spark applications in the cloud

IIT Madras - Research Intern May - June 2016
Model organization of die-stacked DRAM cache for Big Data applications

Selected Projects

Efficient asynchronous RL post-training framework **Under submission**
We design an efficient RL post-training framework that outperforms the fully asynchronous pipeline in verl for multi-turn tool use cases by handling higher off-policy staleness, improving end-to-end throughput by 23% at no accuracy cost. The system also optimizes request scheduling, improving rollout throughput by 30%.

MSCCL++: Rethinking GPU Communication Abstractions for Cutting-edge AI Applications **ASPLOS'26**
We build a fast, flexible, and portable GPU communication library with extensible abstractions that map to different data transfer modes. Using MSCCL++ communication kernels, we speed up collective communication by up to $5.4\times$ and AI inference by up to $1.4\times$. Further, we use MSCCL++ abstractions to build an efficient KV cache transfer engine for prefill-decode disaggregation in LLM inference.

Splitwise: Efficient generative LLM inference using phase splitting **ISCA'24**
We are the first to redesign LLM inference clusters to use different machines for the two inference phases - the prompt computation and token generation phases. Our clusters are optimized for three key objectives: throughput, cost, and power. We can achieve 1.4× higher throughput at 20% lower cost than current designs. Alternatively, we can achieve 2.35× more throughput with the same cost and power budgets.

TACCL: Guiding Collective Algorithm Synthesis using Communication Sketches **NSDI'23**
We build a topology-aware collective communication library to synthesize fast algorithms for collectives like Allgather, Alltoall, and Allreduce commonly used in distributed machine learning. Using the novel concept of

communication sketches provided by a human-in-the-loop, TACCL synthesizes algorithms for NVIDIA DGX-2s that are 25% - $6.7\times$ faster as compared to the state-of-the-art NVIDIA Collective Communication Library (NCCL).

MONeT: Memory Optimizations for Deep Network Training **ICLR'21 spotlight**
We design a checkpointing framework on top of PyTorch which reduces memory requirement of machine learning training by 3x. By jointly optimizing a checkpointing schedule with operator optimizations using a boolean linear programming solver, we minimize the compute overhead of checkpointing to be only about $9-16\%$ of the original PyTorch execution.

Publications

[1] Changho Hwang, Peng Cheng, Roshan Dathathri, Abhinav Jangda, Saeed Maleki, Madan Musuvathi, Olli Saarikivi, **Shah, Aashaka**, Ziyue Yang, Binyang Li, et al. Msccl++: Rethinking gpu communication abstractions for ai inference. In *Proceedings of the 31st ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, pages 1201–1215, 2026.

[2] Pratyush Patel, Esha Choukse, Chaojie Zhang, Íñigo Goiri, **Aashaka Shah**, Saeed Maleki, and Ricardo Bianchini. Splitwise: Efficient generative llm inference using phase splitting. In *51st International Symposium on Computer Architecture (ISCA)*, 2024.

[3] **Aashaka Shah**, Vijay Chidambaram, Meghan Cowan, Saeed Maleki, Madan Musuvathi, Todd Mytkowicz, Jacob Nelson, Olli Saarikivi, and Rachee Singh. Taccl: Guiding collective algorithm synthesis using communication sketches. In *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI)*, 2023.

[4] **Aashaka Shah**, Chao-Yuan Wu, Jayashree Mohan, Vijay Chidambaram, and Philipp Krähenbühl. Memory optimization for deep networks. In *International Conference on Learning Representations (ICLR)*, 2021.

[5] Soujanya Ponnapalli, **Aashaka Shah**, Souvik Banerjee, Dahlia Malkhi, Amy Tai, Vijay Chidambaram, and Michael Wei. Rainblock: Faster transaction processing in public blockchains. In *2021 USENIX Annual Technical Conference (USENIX ATC 21)*, pages 333–347. USENIX Association, July 2021.

[6] **Aashaka Shah**, Vinay Banakar, Supreeth Shastri, Melissa Wasserman, and Vijay Chidambaram. Analyzing the impact of GDPR on storage systems. In *11th USENIX Workshop on Hot Topics in Storage and File Systems (HotStorage 19)*, Renton, WA, 2019. USENIX Association.

[7] Asutosh Palai, Meet Vora, and **Aashaka Shah**. Empowering light nodes in blockchains with block summarization. In *2018 9th IFIP International Conference on New Technologies, Mobility and Security (NTMS)*, pages 1–5, Feb 2018.

Patents

[1] Esha Choukse, Inigo Goiri, Chaojie Zhang, and **Shah, Aashaka**. Heterogenous accelerators for efficient generative llm inference using phase splitting, October 23 2024.

[2] Subrata Mitra, Sopan Khosla, Sanket Vaibhav Mehta, Mekala Rajasekhar Reddy, and **Shah, Aashaka Dhaval**. Tenant-side detection, classification, and mitigation of noisy-neighbor-induced performance degradation, November 21 2019. US Patent App. 15/983,390.

Professional Service

- Served on Program Committees for TPDS 2022, TPDS, 2023, TPDS 2024, PPoPP 2024, ATC 2024, ASPLOS 2025, Eurosys 2026, NeurIPS 2026
- Served as external reviewer for ATC 2025
- Co-organized Women at MSR workshop in 2024
- Served on the Research Poster Committee for SC 2023
- Served as Slack Co-Chair for SOSP 2021
- Served on Eurosys 2020 ShadowPC.
- Co-organized LASR systems seminar at UT Austin for Fall 2018

Invited Talks

- Guest lecture at University of Washington for class: Systems for building foundation models (October 2025)
- Presented MSCCL++ at University of Washington (September 2025)
- Guest lecture at Cornell University for CS 5456: Introduction to Computer Networks (April 2024)
- Presented TACCL at Meta (August 2022)

- Presented MONeT at DeepMind (February 2022)
- Presented "Empowering Light Nodes in Blockchain with Block Summarization" at the India-Japan Workshop on Cryptographic Techniques for Cyber Security, IIT Roorkee (February 2017)

Teaching Experience

- CS360V: Virtualization — Teaching Assistant (UT Austin, Fall 2020)
- CS378H: Concurrency: Honors — Teaching Assistant (UT Austin, Fall 2018)
- CSN-106: Discrete Structures — Teaching Assistant (IIT Roorkee, Spring 2017)