

Rohan Kadekodi

Paul G. Allen School of Computer Science &
Engineering CSE 358
185 E Stevens Way NE, Seattle, WA 98195

🌐 <https://homes.cs.washington.edu/~rohankad>
✉ rohankad@cs.washington.edu
📱 rohankadekodi
☎ (412)-623-9509

Interests

AI Systems, Tiered Memory, Byte Addressable Storage, File and storage systems, Key-value stores, Virtualization, Distributed systems

Education

University of Texas at Austin PhD in Computer Science Advisor: Prof. Vijay Chidambaram	Austin, Texas 2019 - 2023
University of Texas at Austin Master's in Computer Science Advisor: Prof. Vijay Chidambaram	Austin, Texas 2017 - 2019
Pune Institute of Computer Technology Bachelor of Computer Science	Pune, India 2012 - 2016

Work Experience

University of Washington Postdoctoral Scholar Mentor: Prof. Baris Kasikci	Seattle, USA Aug 2023 - Present
Microsoft Research, Redmond Research Intern Mentor: Dr. Badrish Chandramouli	Redmond, USA May 2021 - Aug 2021
Microsoft Research, Redmond Research Intern Mentor: Dr. Badrish Chandramouli	Redmond, USA May 2020 - Aug 2020
Microsoft Research, India Research Intern Mentor: Dr. Harsha Vardhan Simhadri	Bangalore, India May 2019 - Aug 2019
VMware Research Research Intern Mentor: Prof. Aasheesh Kolli	Palo Alto, California May 2018 - Aug 2018
University of Wisconsin, Madison Research Intern Mentors: Prof. Remzi Arpaci-Dusseau and Prof. Vijay Chidambaram	Madison, Wisconsin Jan 2017 - July 2017

Publications

1. Kan Zhu, Tian Tang, Qinyu Xu, Yile Gu, Zhichen Zeng, **Rohan Kadekodi**, Liangyu Zhao, Ang Li, Arvind Krishnamurthy, and Baris Kasikci. Tactic: Adaptive sparse attention with clustering and distribution fitting for long-context llms. In Proceedings of the International Conference on Learning Representations (**ICLR 2026**)
2. Chien-Yu Lin, Keisuke Kamahori, Yiyu Liu, Xiaoxiang Shi, Madhav Kashyap, Yile Gu, Rulin Shao, Zihao Ye, Kan Zhu, **Rohan Kadekodi**, Stephanie Wang, Arvind Krishnamurthy, Luis Ceze, and Baris Kasikci. RagInfer:

Efficient Retrieval-Augmented Generation Inference with Lookahead Retrieval. In Proceedings of the Conference on Machine Learning and Systems (**MLSys 2026**)

3. Yile Gu, Ian Neal, Jiexiao Xu, Shaun Christopher Lee, Ayman Said, Musa Haydar, Jacob Van Geffen, **Rohan Kadekodi**, Andrew Quinn and Baris Kasikci. Scalable and Accurate Application-Level Crash-Consistency Testing via Representative Testing. In Proceedings of the ACM on Programming Languages (**OOPSLA 2025**)
4. Molly Jane Nicholas, Nicolai Marquardt, Michel Pahud, Nathalie Riche, Hugo Romat, Christopher Collins, David Ledo, **Rohan Kadekodi**, Badrish Chandramouli and Ken Hinckley. Escapement: A Tool for Interactive Prototyping with Video via Sensor-Mediated Abstraction of Time. In Proceedings of the CHI Conference on Human Factors in Computing Systems (**CHI 2023**)
5. **Rohan Kadekodi**, Saurabh Kadekodi, Soujanya Ponnappalli, Harshad Shirwadkar, Gregory R. Ganger, Aasheesh Kolli and Vijay Chidambaram. WineFS: a hugepage-aware file system for persistent memory that ages gracefully. In Proceedings of the 28th ACM Symposium on Operating Systems Principles, 2021. (**SOSP 2021**)
6. **Rohan Kadekodi**, Se Kwon Lee, Sanidhya Kashyap, Taesoo Kim, Aasheesh Kolli and Vijay Chidambaram. SplitFS: Reducing Software Overhead in File Systems for Persistent Memory. In Proceedings of the 27th ACM Symposium on Operating Systems Principles, pp. 494-508. ACM, 2019. (**SOSP 2019**)
7. Suhas Jayaram Subramanya, Devvrit, Harsha Simhadri, Ravishankar Krishanswamy, **Rohan Kadekodi**. Rand-NSG: Billion Point Nearest Neighbor Search on a Single Node. Proceedings of the 33rd Annual Conference on Neural Information Processing Systems. (**NeurIPS 2019**)
8. Pandian Raju, **Rohan Kadekodi**, Vijay Chidambaram, Ittai Abraham. PebblesDB: Building Key-Value Stores using Fragmented Log-Structured Merge Trees. Proceedings of the 15th Symposium of Operating Systems Principles, pp. 497-514. ACM, 2017. (**SOSP 2017**)
9. Jayashree Mohan, **Rohan Kadekodi**, Vijay Chidambaram. Analyzing IO Amplification in Linux File Systems (Poster). Proceedings of the 8th ACM The Eighth SIGOPS AsiaPacific Workshop on Systems, Sep 2017. **Best Poster Award**. (**ApSys 2017**)

Teaching

- | | |
|--|--|
| 1. CS360V: Virtualization
Assistant Instructor, UT Austin | Austin, Texas
<i>Spring 2023</i> |
| 2. CS380D: Distributed Computing
Teaching Assistant, UT Austin | Austin, Texas
<i>Spring 2022</i> |
| 3. CS378: Virtualization
Teaching Assistant, UT Austin | Austin, Texas
<i>Fall 2019</i> |

Technical Talks

1. Profile Guided Memory Tiering (PRISM SRC Liaison Talk). May 2024.
2. Building high-performance storage systems for Persistent Memory (BigHPC Webinar). July 2022.
3. WineFS: A hugepage-aware file system for persistent memory that ages gracefully at Storage Analytics team (Google). October 2021.
4. WineFS: A hugepage-aware file system for persistent memory that ages gracefully at Symposium on Operating Systems Principles (SOSP 2021). October 2021.
5. SplitFS: Reducing Software Overhead in File Systems for Persistent Memory at Symposium on Operating Systems Principles (SOSP 2019), Huntsville, Canada. October 2019.

Academic Projects

Fine-tuning local LLMs for agentic workflows

University of Washington (Ongoing)

Exploring fine-tuning techniques for local LLMs to improve their tool calling capabilities in agentic systems.

Benchmarking and Improving the Performance of LLM Applications on End-User Devices

University of Washington (Ongoing)

Building a benchmarking suite and runtime system for accelerating the performance of agentic workflows for end-user devices such as laptops and mobile phones, using smart GPU sharing and memory management.

Tiered Memory Management with Controlled Allocation and Adaptive Migration

University of Washington (Ongoing)

Building a tiered memory system that uses smart allocation policies for achieving fine-grained hot data tracking, along with adaptive migrations for reacting to workload changes.

RDFS: Enabling Remote DAX Memory-Mapping for Persistent Memory

University of Washington & UT Austin (Ongoing)

Building a distributed Persistent Memory manager for transparent scaling of single-node memory-mapped applications across a cluster.

Shared-state system for distributed interactive applications

Microsoft Research (CHI 2023)

Building a shared-state system with client-side caching and easy-to-use API for supporting distributed interactive applications in the cloud.

WineFS: a hugepage-aware file system for persistent memory that ages gracefully

UT Austin (SOSP 2021)

A PM file system aimed at preserving hugepages for improving the performance of emerging PM applications.

Scaling Nearest Neighbor Search to a single-node secondary storage and distributed cluster

Microsoft Research (NeurIPS 2019)

Scaling Rand-NSG approximate Nearest Neighbor to out-of-core storage on a single node and a distributed cluster, while maintaining performance and availability.

SplitFS: Reducing Software Overhead in File Systems for Persistent Memory

UT Austin & VMware Research (SOSP 2019)

A user-space file system aimed at improving performance of POSIX applications on persistent memory by converting reads and writes to loads and stores from user space, and passing metadata operations to the kernel.

Sequentiality Matters! Even on Persistent Memory

UT Austin

We Measure the effect of sequential and random access patterns on byte-addressable Persistent Memory due to CPU caches and TLB, and conclude that sequential accesses significantly outperform random accesses.

Analyzing IO Amplification in Linux File Systems

UT Austin (Best Poster at ApSys 2017)

We present an analysis the IO amplification of ext4, btrfs, xfs and f2fs for different real-world workloads and introduce a new principled approach to design file systems.

PebblesDB: Building Key-Value Stores using Fragmented Log-Structured Merge Trees

UT Austin (SOSP 2017)

A key-value store based on fragmented log-structured merge trees, which reduces IO Amplification while increasing throughput.

Professional Service

Program Committee Member

- Non-volatile Memories Workshop (NVMW) - 2024
- Annual Technical Conference (ATC) - 2024, 2025
- Operating Systems Design & Implementation (OSDI) - 2025

Awards and Achievements

- Best Poster Award, ApSys 2017
- Winner, Purushottam Karandak 2016
- Winner, Best Project, Quarks 2016
- Winner, Rolocule Innovation, Impetus and Concepts 2016
- Runner-up, Best Project, Impetus and Concepts 2016