# ANURAG KHANDELWAL

**ASSISTANT PROFESSOR, YALE UNIVERSITY**

**EMAIL:** anurag.khandelwal@yale.edu **WEBSITE:** http://anuragkhandelwal.com


## Education

**University of California Berkeley**
*PhD in Computer Science (2013 - 2019), Advised by Prof. Ion Stoica*

**Indian Institute of Technology, Kharagpur**
*B. Tech. in Computer Science and Engineering (2009 - 2013)*

## Work History

**Yale University, New Haven**
*Assistant Professor (tenure-track), Spring 2020 - Present*

**Cornell University, New York Campus**
*Postdoctoral Associate, Fall 2019*

**University of California, Berkeley**
*Graduate Student Researcher, Fall 2013 - Summer 2019*

**Microsoft Research**
*Summer Research Intern, 2014 and 2012*

## Awards & Honors

**Distinguished Artifact Award, ACM ISCA 2025**
For paper "*CORD: Low-Latency, Bandwidth-Efficient and Scalable Release Consistency via Directory Ordering*"

**NetApp Faculty Fellowship, 2024**
To develop GPU-native vector storage and indexing for accelerating AI workloads.

**Best Student Paper Award, ACM EuroSys 2024**
For paper "*TRINITY: A Fast Compressed Multi-attribute Data Store*"

**IEEE MICRO's Top-picks in Computer Architecture, 2024**
For paper "*SCALO: An Accelerator-Rich Distributed System for Scalable Brain-Computer Interfaces*"

**Best Paper Award, ACM ISCA 2023**
For paper "*SCALO: An Accelerator-Rich Distributed System for Scalable Brain-Computer Interfaces*"

**Roberts Innovation Award, 2023**
For innovations in data center resource disaggregation.

**NSF CAREER Award, 2021**
For developing in-network memory management for disaggregated data centers.

**NetApp Faculty Fellowship, 2021**
For developing elastic and compressed disaggregated memory for edge data centers.

**Distinguished Paper Award, USENIX Security 2020**
For paper "*Pancake: Frequency Smoothing for Encrypted Data Stores*"

**Madan Sundar Sahu Scholarship 2010, 2011, Indian Institute of Technology, Kharagpur**
For outstanding academic performance during 2009-2010 and 2010-2011.

## Research Focus

*My research interests span systems, networks, and systems security.*
My work addresses challenges in processing, storing, and serving large volumes of data to empower real-world systems: from sprawling internet services like social media to critical tools in health and medicine. My research formulates such challenges as algorithm and data structure design problems, resulting in practical systems backed by strong theoretical guarantees that have found their way into production.

## Research Impact

[2025] CORD is influencing the design of future NVIDIA GPU-CPU architectures.

[2024] PromptCache adopted in Gemini, OpenAI, and Anthropic for reusing attention states across LLM prompts.

[2024] Trinity is being integrated into NetApp's filesystem services to improve storage efficiency and performance.

[2023] The ideas underlying SCALO are being tested by Yale Neurosurgery in FPGA-based emulation frameworks for interfacing with the human brain in the operating room.

[2021] My NSF CAREER Award project, MIND, is the foundational system and central pillar of the multi-institution $20-million NSF-funded AI Institute, Athena.

[2019] Confluo employed in Alibaba Cloud for consuming and analyzing telemetry data.

[2015] Succinct employed by Databricks customers to query compressed RDDs.

## Publications

My research publications span two key areas of computer science: (1) *systems & networks* and (2) *systems security*. Research in both communities (and, indeed, computer science in general) tend to focus on publications at conferences and workshops over journals[1]. As such, my own publications span top-tier conferences in both classes. *Students/postdocs under my supervision are highlighted in bold in the author list.*

### Conferences & Workshops

[C1] CounterPoint: Using Hardware Event Counters to Refute and Refine Microarchitectural Assumptions. Nick Lindsay, Caroline Trippel, **Anurag Khandelwal**, and Abhishek Bhattacharjee. In *ACM ASPLOS*, 2026

[C2] Spirit: Fair Allocation of Interdependent Resources in Remote Memory Systems. **Seung-seob Lee**, **Jachym Putta**, **Ziming Mao**, and **Anurag Khandelwal**. In *ACM SOSP*, 2025

[C3] Scalable Far Memory: Balancing Faults and Evictions. Yueyang Pan[*], **Yash Lala**[*], Musa Unal, Yujie Ren, **Seung-seob Lee**, Abhishek Bhattacharjee, **Anurag Khandelwal**, and Sanidhya Kashyap. In *ACM SOSP*, 2025. ([*]Equal contribution authors)

[C4] Found in Translation: A Generative Language Modeling Approach to Memory Access Pattern Attacks. **Grace Jia**, Alex Wong, and **Anurag Khandelwal**. In *USENIX Security*, 2025

[C5] Weave: Efficient and Expressive Oblivious Analytics at Scale. **Mahdi Soleimani**, **Grace Jia**, and **Anurag Khandelwal**. In *USENIX OSDI*, 2025

[C6] CORD: Low-Latency, Bandwidth-Efficient and Scalable Release Consistency via Directory Ordering. **Yanpeng Yu**, Nicolai Oswald, and **Anurag Khandelwal**. In *ACM ISCA*, 2025. [Distinguished Artifact Award]

[C7] PULSE: Accelerating Distributed Pointer-Traversals on Disaggregated Memory. **Yupeng Tang**, **Seung-seob Lee**, Abhishek Bhattacharjee, and **Anurag Khandelwal**. In *ACM ASPLOS*, 2025

[C8] Exploring Intelligent Dynamic Resource Provisioning for Elastic Massive MIMO vRAN. Parthiban Annamalai, Minsung Kim, **Anurag Khandelwal**, and Lin Zhong. In *ACM HotMobile*, 2025

[C9] Length Leakage in Oblivious Data Access Mechanisms. **Grace Jia**, Rachit Agarwal, and **Anurag Khandelwal**. In *USENIX Security*, 2024

[C10] Distributed Brain–Computer Interfacing With a Networked Multiaccelerator Architecture. Raghavendra Pothukuchi, **Karthik Sriram**, Michał Gerasimiuk, Muhammed Ugur, Rajit Manohar, **Anurag Khandelwal**, and Abhishek Bhattacharjee. *IEEE Micro*, 2024

[C11] Prompt Cache: Modular attention reuse for low-latency inference. In Gim, Guojun Chen, **Seung-seob Lee**, Nikhil Sarda, **Anurag Khandelwal**, and Lin Zhong. *MLSys*, 2024

[C12] Trinity: A Fast Compressed Multi-attribute Data Store. **Ziming Mao**, Kiran Srinivasan, and **Anurag Khandelwal**. In *ACM EuroSys*, 2024. [Best Student Paper Award]

[C13] Karma: Resource Allocation for Dynamic Demands. Midhul Vuppalapati, Giannis Fikioris, Rachit Agarwal, Asaf Cidon, **Anurag Khandelwal**, and Éva Tardos. In *USENIX OSDI*, 2023

[C14] Prefetching Using Principles of Hippocampal-Neocortical Interaction. **Michael Wu**, Ketaki Joshi, Andrew Sheinberg, Guilherme Cox, **Anurag Khandelwal**, Raghavendra Pradyumna Pothukuchi, and Abhishek Bhattacharjee. In *ACM HotOS*, 2023

[C15] SCALO: An Accelerator-Rich Distributed System for Scalable Brain-Computer Interfacing. **Karthik Sriram**, Raghavendra Pothukuchi, Michał Gerasimiuk, Muhammed Ugur, Oliver Ye, Rajit Manohar, **Anurag Khandelwal**, and Abhishek Bhattacharjee. In *ACM ISCA*, 2023. [Best Paper Award][Selected for inclusion in IEEE Micro's Top Picks in Computer Architecture]

[C16] Shepherd: Serving DNNs in the Wild. **Hong Zhang**, **Yupeng Tang**, **Anurag Khandelwal**, and Ion Stoica. In *USENIX NSDI*, 2023

[C17] Shortstack: Distributed, Fault-tolerant, Oblivious Data Access. Midhul Vuppalapati[*], Kushal Babel[*], **Anurag Khandelwal**, and Rachit Agarwal. In *USENIX OSDI*, 2022. ([*]Equal contribution authors)

[C18] Jiffy: Elastic Far-memory for Stateful Serverless Analytics. **Anurag Khandelwal**, **Yupeng Tang**, Rachit Agarwal, Aditya Akella, and Ion Stoica. In *ACM EuroSys*, 2022

[C19] MIND: In-Network Memory Management for Disaggregated Data Centers. **Seung-seob Lee**, **Yanpeng Yu**, **Yupeng Tang**, **Anurag Khandelwal**, Lin Zhong, and Abhishek Bhattacharjee. In *ACM SOSP*, 2021

[C20] What Serverless Computing Is and Should Become: The Next Phase of Cloud Computing. Johann Schleier-Smith, Vikram Sreekanti, **Anurag Khandelwal**, Joao Carreira, Neeraja J. Yadwadkar, Raluca Ada Popa, Joseph E. Gonzalez, Ion Stoica, and David A. Patterson. *Communications of the ACM*, 2021

[C21] Caerus: Nimble Task Scheduling for Serverless Analytics. **Hong Zhang**, **Yupeng Tang**, **Anurag Khandelwal**, Jingrong Chen, and Ion Stoica. In *USENIX NSDI*, 2021

[C22] Pancake: Frequency Smoothing for Encrypted Data Stores. Paul Grubbs[*], **Anurag Khandelwal**[*], Marie-Sarah Lacharité[*], Lloyd Brown, Lucy Li, Rachit Agarwal, and Thomas Ristenpart. In *USENIX Security*, 2020. [Distinguished Paper Award] ([*]Equal contribution authors)

[C23] Le Taureau: Deconstructing the Serverless Landscape & A Look Forward. **Anurag Khandelwal**, Arun Kejariwal, and Karthikeyan Ramasamy. In *ACM SIGMOD*, 2020

[C24] Confluo: Distributed Monitoring and Diagnosis Stack for High-speed Networks. **Anurag Khandelwal**, Rachit Agarwal, and Ion Stoica. In *USENIX NSDI*, 2019

[C25] ZipG: A Memory-efficient Graph Store for Interactive Queries. **Anurag Khandelwal**, Zongheng Yang, Evan Ye, Rachit Agarwal, and Ion Stoica. In *ACM SIGMOD*, 2017

[C26] BlowFish: Dynamic Storage-Performance Tradeoff in Data Stores. **Anurag Khandelwal**, Rachit Agarwal, and Ion Stoica. In *USENIX NSDI*, 2016

[C27] Succinct: Enabling Queries on Compressed Data. Rachit Agarwal, **Anurag Khandelwal**, and Ion Stoica. In *USENIX NSDI*, 2015

## Theses

[T1] *Queries on compressed data*. **Anurag Khandelwal**. University of California, Berkeley, 2019

## Technical Reports

[R1] Cloud Programming Simplified: A Berkeley View on Serverless Computing. Eric Jonas, Johann Schleier-Smith, Vikram Sreekanti, Chia-Che Tsai, **Anurag Khandelwal**, Qifan Pu, Vaishaal Shankar, João Carreira, Karl Krauth, Neeraja J. Yadwadkar, Joseph E. Gonzalez, Raluca Ada Popa, Ion Stoica, and David A. Patterson. *UC Berkeley Technical Report*, 2019

[R2] Attacking Data Center Networks from the Inside. **Anurag Khandelwal**, Navendu Jain, and Seny Kamara. *Microsoft Research Technical Report*, 2015

## Preprints

[P1] Efficient MoE Serving in the Memory-Bound Regime: Balance Activated Experts, Not Tokens. **Yanpeng Yu**, Haiyue Ma, Krish Agarwal, Nicolai Oswald, Qijing Huang, Hugo Linsenmaier, Chunhui Mei, Ritchie Zhao, Ritika Borkar, Bita Darvish Rouhani, David Nellans, Ronny Krashinsky, and **Anurag Khandelwal**. *arXiV Paper 2512/09277*, 2025

[P2] Wiretapping LLMs: Network Side-Channel Attacks on Interactive LLM Services. **Mahdi Soleimani**, **Grace Jia**, In Gim, **Seung-seob Lee**, and **Anurag Khandelwal**. *Cryptology ePrint Archive, Paper 2025/167*, 2025

[P3] GCS: Generalized Cache Coherence For Efficient Synchronization. **Yanpeng Yu**, **Seung-seob Lee**, Lin Zhong, and **Anurag Khandelwal**. *arXiv Paper 2301/02576*, 2023

## Invited Talks

Performant Shared Disaggregated Memory
*Carnegie Mellon University, November 2025*
*University of Washington, October 2025*
*University of Texas–Austin, October 2025*
*Princeton University, October 2025*
*University of Pennsylvania, October 2025*
*Massachusetts Institute of Technology, September 2025*
*New York University, March 2025*
*Brown University, March 2025*
*NSF AI Institute Showcase, Duke University, August 2024*
*Microsoft Research, August 2024*
*AMD Research, April 2024*

In-Network Memory Management for Disaggregated Data Centers
*NSF AI Institute Showcase, Duke University, August 2022*
*Cornell University, November 2021*
*VMWare, November 2021*
*Columbia University, October 2021*

Jiffy: Ephemeral Storage for Serverless Analytics
*Facebook, March 2021*
*Intel, March 2021*

Serverless streaming architectures & algorithms for the enterprise
*Global STAC Live, October 2020 (Industry conference for industry leaders in technology.)*
*Strata Data Conference, September 2019 (Industry conference of over 5000 companies.)*

Building Interactive Query Systems
*Yale University, April 2019*
*Purdue University, March 2019*
*Microsoft Research, Redmond, March 2019*
*Microsoft Research, Cambridge (UK), March 2019*

Realtime Monitoring and Analysis with Confluo
*Splunk, San Francisco, February 2018*

Succinct: Enabling Queries on Compressed Data
*SF Big Analytics Meetup, San Francisco, September 2016*
*Bay Area AI Meetup, San Francisco, September 2016*
*Cloudera, San Fransisco, June 2016*

## Funding

NSF SaTC: CORE: Medium: Fortifying and Enriching Confidential Computing Environments
2025-28, $900,000; PI: **Anurag Khandelwal**, Co-PIs: Lin Zhong, Seung-seob Lee

NetApp Faculty Fellowship: AI-Native Vector Storage and Indexing
2024-25, $117,686; PI: **Anurag Khandelwal**

Yale Engineering AI Seed Grant: Brain-Inspired Memory Systems for AI Infrastructure
2024-25, $150,000; PI: Abhishek Bhattacharjee, Co-PI: **Anurag Khandelwal**

Roberts Innovation Fund: Disaggregated Edge Clouds
2023-24, $50,000; PI: **Anurag Khandelwal**

NSF RINGS: Intelligent and Resilient Virtualization of Massive MIMO Physical Layer
2022-25, $1,000,000; PI: Lin Zhong, Co-PI: **Anurag Khandelwal**

Swebilius Foundation Award
2022, $16,000; PI: Abhishek Bhattacharjee, Co-PI: **Anurag Khandelwal**, Rajit Manohar

NSF AI Institute for Edge Computing Leveraging Next Generation Networks
2021-26, $20,000,000; Multi-institution, multi-PI grant, $575,000 to **Khandelwal**

NSF PPoSS: Planning: High-Performance Certified Trust for Global-Scale Applications
2021-22, $250,000; PI: Zhong Shao, Co-PIs: Abhishek Bhattacharjee, **Anurag Khandelwal**, Lin Zhong

NSF SaTC: CORE: Medium: Mixed Distribution Models for Encrypted Data Stores
2021-25, $1,000,000; PI: **Anurag Khandelwal**, Co-PIs: Rachit Agarwal, Thomas Ristenpart

NetApp Faculty Fellowship: A Unified Data Stack for Next-Generation IoT Systems
2021, $50,000; PI: **Anurag Khandelwal**, Co-PI: Abhishek Bhattacharjee

NSF CAREER: In-Network Memory Management for Disaggregated Datacenters
2021-26, $626,647.00; PI: **Anurag Khandelwal**

## Teaching

CPSC422/522: Operating Systems
- Spring 2025: Course Director, 24 Lectures, Enrollment: 34 (YC)
- Spring 2024: Course Director, 24 Lectures, Enrollment: 44 (YC)

CPSC433/533: Computer Networks
- Spring 2022: Course Director, 24 Lectures, Enrollment: 39 (YC)
- Spring 2021: Course Director, 24 Lectures, Enrollment: 34 (YC)
- Spring 2020: Course Director, 24 Lectures, Enrollment: 44 (YC)

CPSC 438/538: Big Data Systems
- Fall 2023: Course Director, 26 Lectures, Enrollment: 18 (YC)
- Fall 2021: Course Director, 26 Lectures, Enrollment: 16 (YC)

CPSC 638: Big Data Systems
- Fall 2020: Course Director, 26 Lectures, Enrollment: 3 (YC)

## Advising

Postdoctoral Research:
Seung-seob Lee (2020—, co-advised with Lin Zhong)

Graduate Student Research:
Yupeng Tang (2020—2024, PhD) (→ Research Scientist at Meta Research)
Yanpeng Yu (2021—)
Grace Jia (2022—)
Yash Lala (2023—)
Mahdi Solemani (2023—)
Jachym Putta (2023—2025; Masters)

Undergraduate Student Research: Michael Tu (2024), Troy Feng (2023), Raymond Yang (2023), Ziming Mao (2021—24), Gabriel Buchdahl (2022—23), Jachym Putta (2022—23), Adit Gupta (2022—23), Yuxuan Chen (2022), Eli Sage-Martinson (2022), Stanley Wong (2022), Daniel Li (2021—22), Brian Liu (2021), Alan Chen (2021), Jonathan Kraft (2021), Eric Chang (2020)

## Education Outreach

Yale Pathways to Science, Summer Scholars Program
- *"Through the Clouds: An Introduction to Cloud Computing"*
- Instructor for 4-day workshop for high-school students (Summer 2021)
- Instructor for 1-day seminar for high-school students (Summer 2022)

## Community Service

Organization Committee: CoNEXT'23 Poster Co-Chair
Program Committee: SIGCOMM 2020, NSDI 2021, NSDI 2022, HotNets 2022, NSDI 2023, EuroSys 2023, SOSP 2024, EuroSys 2025, OSDI 2025, NSDI 2026
Journal Editorial Board (Serverless Area): JSys 2021
External Program Committee: ASPLOS 2021
NSF Panels: Undisclosed due to participant confidentiality.

## Department Service

Chair, Colloquium Committee: Fall 2021
CS Undergraduate Advisor: 2021-22 (Freshman/Sophomore), 2024-25 (Junior), 2025-26 (Junior)
Yale CS Research Internship Program: Created and led the program (2021-25)
Faculty Hiring Committee: 2025 (Database Systems Area Search)
PhD Admissions Committee: 2020, 2021, 2022, 2023, 2024, 2025
Summer Systems Seminars: 2020, 2021, 2022
PhD Thesis Committee: Bo Hu (2021), Karthik Sriram (2023), Yupeng Tang (2024)