Juncheng Yang

☎ (+1) 404-285-5231 | ✉ juncheng@seas.harvard.edu | ⌂ http://junchengyang.com

Academic Positions

Assistant Professor @ Harvard University — *Cambridge*
School of Engineering and Applied Science — *July 2025 - Present*

Industry Positions

Research Scientist @ Snowflake — *Remote*
Snowflake AI research, manager: Yuxiong He — *Mar 2025 - June 2025*

Postdoctoral Scientist @ AWS — *Boston*
S3, manager: James Bornholt — *Sept 2024 - Mar 2025*

Software Engineer Intern @ Twitter — *San Francisco*
JVM off-heap caching, manager: Yao Yue — *May 2022 - July 2022*

Software Engineer Intern @ Cloudflare — *Remote*
Content delivery network performance, manager: Aki Shugaeva — *June 2021 - Aug 2021*

Researcher @ Twitter — *Remote*
In-memory caching, manager: Yao Yue — *Feb 2020 - Nov 2020*

Education

Ph.D. in Computer Science, Carnegie Mellon University — *Pittsburgh, U.S.A*
Computer Science Department, advisor: Rashmi Vinayak — *Aug. 2018 - Sept 2024*

M.S. in Computer Science, Emory University — *Atlanta, U.S.A*
Department of Mathematics and Computer Science, advisor: Ymir Vigfusson — *Jan. 2015 - Dec. 2016*

M.S. in Chemistry, Emory University — *Atlanta, U.S.A*
Department of Chemistry, advisor: Craig L. Hill — *Aug. 2013 - Jun. 2015*

B.S. in Chemistry, Nanjing University — *Nanjing, China*
Department of Chemistry and Chemical Engineering, advisor: Ying Wang — *Sept. 2009 - Jun. 2013*

Research Highlights

SIEVE — The ultimate cache eviction algorithm that is simpler than LRU with state-of-the-art efficiency and scalability. Implemented and deployed at over 20 companies and open-source libraries in more than 16 programming languages. Community award at NSDI'24. Find more at `https://sieve-cache.com`.

S3-FIFO — A simple and scalable cache eviction algorithm, implemented or deployed at companies including Google, AWS, VMware, Redpanda, and many others, with many open-source libraries. Find more at `https://s3fifo.com`.

Segcache — A new storage layout for modern key-value caches. Received a community award at NSDI'21, deployed at Twitter and Momento.

Selected Honors & Awards

2024	**NSDI’24 Community (Best Paper) Award**
2023	**Machine Learning and System Rising Star**
2023	**Google Cloud Research Innovator**
2020-2022	**Meta Fellowship**
2021	**SOSP’21 Best Paper Award**
2021	**NSDI’21 Community (Best Paper) Award**
2016	**SYSTOR’16 Best Student Paper**
2012	**“Person of the Year” Nomination** 100 nominations among all Chinese undergraduates.

Selected Publications

Machine learning and system

ASPLOS’25 Yixuan Mei, Yonghao Zhuang, Xupeng Miao, Juncheng Yang, Zhihao Jia, K. V. Rashmi. **“Helium: Serving Large Language Models on Heterogeneous GPUs via Max-Flow.”** *the ACM International Conference on Architectural Support for Programming Languages and Operating Systems*.

FAST’23 Juncheng Yang, Ziming Mao, Yao Yue, K. V. Rashmi. **“GL-Cache: Group-level learning for efficient and high-performance caching.”** *The 21st USENIX Conference on File and Storage Technologies*.

SOCC’17 Juncheng Yang, Reza Karimi, Trausti Saemundsson, Avani Wildani, Ymir Vigfusson. **“Mithril Mining Sporadic Associations for Cache Prefetching.”** *ACM Symposium on Cloud Computing*.

VLDB’23 Tianyu Zhang, Kaige Liu, Jack Kosaian, Juncheng Yang, K. V. Rashmi. **“Efficient Fault Tolerance for Recommendation Model Training via Erasure Coding.”** *49th International Conference on Very Large Database*.

Storage system and database

NSDI’24 Yazhuo Zhang* (mentored student), Juncheng Yang*, Yao Yue, Ymir Vigfusson, K. V. Rashmi. **“SIEVE is Simpler than LRU: an Efficient Turn-Key Eviction Algorithm for Web Caches.”** *The 21st USENIX Symposium on Networked System Design and Implementation*. **Community (Best Paper) Award**.

SOSP’23 Juncheng Yang, Yazhuo Zhang, Ziyue Qiu, Yao Yue, K. V. Rashmi. **“FIFO Queues are All You Need for Cache Eviction.”** *ACM Symposium on Operating System Principles*.

HotOS’23 Juncheng Yang, Ziyue Qiu, Yazhuo Zhang, Yao Yue, K. V. Rashmi. **“FIFO Can be Better than LRU: the Power of Lazy Promotion and Quick Demotion.”** *The 19th Workshop on Hot Topics in Operating Systems*.

SOCC’23 Yazhuo Zhang, Rebecca Isaacs, Yao Yue, Juncheng Yang, Lei Zhang, Ymir Vigfusson. **“Latenseer: Causal Modeling of End-to-End Latency Distributions by Harnessing Distributed Tracing.”** *ACM Symposium on Cloud Computing*.

Eurosys’23 Ziyue Qiu, Juncheng Yang, Juncheng Zhang, Cheng Li, Xiaosong Ma, Qi Chen, Mao Yang, Yinlong Xu. **“FrozenHot Cache: Rethinking Cache Management for Modern Hardware.”** *The European Conference on Computer Systems*.

NSDI’22 Juncheng Yang, Anirudh Sabnis, Daniel S. Berger, K. V. Rashmi, Ramesh K. Sitaraman. **“C2DN: How to Harness Erasure Codes at the Edge for Efficient Content Delivery.”** *19th USENIX Symposium on Networked Systems Design and Implementation*.

NSDI’21 Juncheng Yang, Yao Yue, K. V. Rashmi. **“Segcache: memory-efficient and high-throughput DRAM cache for small objects.”** *18th USENIX Symposium on Networked Systems Design and Implementation*. **Community (Best Paper) Award**.

SOSP’21 Sara McAllister, Benjamin Berg, Julian Tutuncu-Macias, Juncheng Yang, Sathya Gunasekar, Jimmy Lu, Nathan Beckmann, Gregory R. Ganger. **“Kangaroo: Caching Billions of Tiny Objects on Flash.”** *28th ACM Symposium on Operating Systems Principles*. **Best Paper Award, invited fast-track to TOS’22**.

OSDI’20 Juncheng Yang, Yao Yue, K. V. Rashmi. **“A Large Scale Analysis of Hundreds of In-memory Cache Clusters at Twitter.”** *14th USENIX Symposium on Operating Systems Design and Implementation*. **Invited fast track submission to TOS’21**.

OSDI’20 Saurabh Kadekodi, Francisco Maturana, Suhas Jayaram Subramanya, Juncheng Yang, K. V. Rashmi, Gregory R. Ganger. **“PACEMAKER: Avoiding HeART Attacks in Storage Clusters with Disk-adaptive Redundancy.”** *14th USENIX Symposium on Operating Systems Design and Implementation*.

SOCC’18 Hobin Yoon, Juncheng Yang, Sveinn Fannar Kristjansson, Steinn E. Sigurdarson, Ymir Vigfusson, Ada Gavrilovska. **“Mutant: Balancing Storage Cost and Latency in LSM-Tree Data Stores.”** *ACM Symposium on Cloud Computing*.

ICDE’18 Jinfei Liu, Juncheng Yang, Li Xiong, Jian Pei, Jun Luo. **“Skyline Diagram: Finding the Voronoi Counterpart for Skyline Queries.”** *IEEE International Conference on Data Engineering*.

ICDE’17 Jinfei Liu, Juncheng Yang, Li Xiong, Jian Pei. **“Secure Skyline Queries on Cloud Platform.”** *IEEE International Conference on Data Engineering*.

SYSTOR’16 Helgi Sigurbjarnarson, Petur Orri Ragnarsson, Juncheng Yang, Ymir Vigfusson, Mahesh Balakrishnan. **“Enabling Space Elasticity in Storage Systems.”** *ACM International Systems and Storage Conference*. **Best Student Paper Award**.

Invited Talk

1. FIFO queues are all you need for cache eviction.
   - Workshop on Streaming (WOS’23), 2023
   - VMware, 2023
   - Alluxio, 2023
   - Microsoft Research Asia, 2023
   - Kuaishou, 2023
   - University of Science and Technology of China, 2023
   - Tsinghua University, 2023
2. LESSCache: LEarned Segment-Structured cache.
   - Meta, 2023
   - VMware, 2022
3. Ubiquitous caching: building efficient distributed and in-process caching. *QCon SF*, 2022.
4. Segcache: a memory-efficient and high-throughput DRAM cache for small objects.
   - Oracle, 2023
   - Alluxio, 2022
5. Caching on PMEM: an iterative approach. *SNIA SDC keynote talk*, 2020.

Funding and grants

2023	**Google Cloud Innovator grant**	$20,000
2018	**AWS research grant**	$10,000

Open Source Contributions

2018-2025	**libCacheSim**	A high-performance cache simulator	*Carnegie Mellon University*
2020-2025	**distComp**	A fault-tolerant and memory-adaptive distributed computation platform	*Carnegie Mellon University*
2021-2023	**fastscp**	A fast data transfer tool using CDN overlay network	*Carnegie Mellon University*
2020-2021	**Segcache**	A prototype of segment-structured cache	*Carnegie Mellon University*
2016-2018	**mimircache**	A Python package for cache performance analysis and visualization	*Emory University*

Service & Activities

External Service

2025-2026 **Artifact evaluation chair for FAST’25, FAST’26**

2024-2026 **Program Committee for SOSP’24 Poster, FAST’25, ICDCS’25, mlsys’25, FAST’26**

2022-2026 **Journal Reviewer for IEEE TKDE, TMC, SC, TCC, TPDS, Access, ACM TOS**

Internal Service

2023-2024	**Organizer** Parallel Data Lab reading group
2023	**Organizer** Parallel Data Lab retreat practice talk series
2020-2023	**Organizer** CMU school of computer science student speaking seminar series

Teaching

2022 & 2023	**Guest lecturer** 15612 Intro to Computer System	*Carnegie Mellon University*
2022	**Teaching assistant** 15712 Advanced and Distributed Operating Systems	*Carnegie Mellon University*
2020	**Teaching assistant** 15746 Storage Systems	*Carnegie Mellon University*
2017	**Guest lecturer** CS584 Advanced Computer System	*Emory University*
2017	**Teaching assistant** CS453 Computer Security	*Emory University*
2013, 2014	**Lab instructor** General Chemistry I and II	*Emory University*
2012	**Teaching assistant** Modern Website Programming	*Nanjing University*

Mentees

2021-2023	Jonathan Chiu (CMU undergraduate)
2022	Ziming Mao (Yale undergraduate, UC Berkeley Ph.D.)
2022-2024	Yazhuo Zhang (Emory Ph.D.)
2022-2025	Ziyue Qiu (CMU Ph.D.)
2023	Emily Zhang (CMU undergraduate)
2023	Parinay Chauhan (IIT undergraduate)
2023-2024	Frank Chen (CMU undergraduate)
2024	Helen Wang (CMU undergraduate)
2023-2025	Bob Chen (CMU undergraduate)
2023-2025	Yiyan Zhai (CMU undergraduate)
2025	Hongshu Yan (ETHz master)
2025	Bintang Dwi Marthen (ITB undergraduate)
2025	Raden Rafly H. B. (ITB undergraduate)
2025	Muhammad Haekal M. A. (ITS undergraduate)
2025	Mingyan Gao (ZJU and UIUC undergraduate)