

Danyang Zhuo

Assistant Professor

Department of Computer Science
Trinity College of Arts and Sciences
Duke University

November 8, 2025

308 Research Dr

Durham, NC 27705

danyang@cs.duke.edu

<https://danyangzhuo.com>

Education

- **University of California – Berkeley** Berkeley, California
Postdoctoral Researcher *Sep 2019 - Jun 2020*
 - Advisor: Ion Stoica
- **University of Washington – Seattle** Seattle, Washington
Ph.D. in Computer Science and Engineering *Sep 2013 - Aug 2019*
 - Dissertation: Practical, Efficient, and Reliable Data Center Communication.
 - Advisors: Thomas E. Anderson, Arvind Krishnamurthy
- **University of Illinois – Urbana Champaign** Urbana, Illinois
B.S. in Electrical Engineering *Aug 2009 - May 2013*
 - Advisor: Nitin Vaidya

Professional Experience

- **Duke University** Durham, North Carolina
Assistant Professor of Computer Science *Jul 2020 - now*
- **Apple** Remote
Visiting Researcher (through Magnit) *Feb 2025 - now*
- **Microsoft Research** Redmond, Washington
Contractor (through Populous Group) *Oct 2015 - Feb 2017*
- **Microsoft Research** Redmond, Washington
Research Intern *Jun 2015 - Sep 2015*
- **Google** Mountain View, California
Software Development Engineering Intern *Sep 2014 - Mar 2015*
- **Amazon** Seattle, Washington
Software Development Engineering Intern *May 2013 - Sep 2013*
- **Microsoft** Redmond, Washington
Software Development Engineering Intern *May 2012 - Aug 2012*

Awards

- IEEE/ACM MICRO Best Paper Award 2025
- Google Academic Research Award 2024
- ACSIC Rising Star Award 2024

- NSF CAREER Award 2023
- USENIX Security Distinguished Paper Award 2023
- Meta Research Award 2022
- USENIX FAST Best Paper Award 2021
- Amazon Research Award 2021
- IBM Academic Award 2021
- Meta Research Award 2021
- University of Washington Madrona Prize Runner-Up 2018
- University of Washington Hacherl Endowed Fellowship 2013
- Rank 146th in the William Lowell Putnam Mathematical Competition 2012

Publications

Conference Papers

1. Zishan Shao, Yixiao Wang, Qinsi Wang, Ting Jiang, Zhixu Du, Hancheng Ye, Danyang Zhuo, Yiran Chen, Hai Helen Li. *FlashSVD: Memory-Efficient Inference with Streaming for Low-Rank Models*. The 40th Annual AAAI Conference on Artificial Intelligence (AAAI), 2026.
2. Xiangfeng Zhu, Yuyao Wang, Banruo Liu, Yongtong Wu, Nikola Bojanic, Jingrong Chen, Gilbert Bernstein, Arvind Krishnamurthy, Sam Kumar, Ratul Mahajan, Danyang Zhuo. *High-level Programming for Application Networks*. The 22th USENIX Symposium on Networked Systems Design and Implementation (NSDI), 2025.
3. Ceyu Xu, Yongji Wu, Xinyu Yang, Beidi Chen, Matthew Lentz, Danyang Zhuo, Lisa Wu Wills. *LLM.265: Video Codecs are Secretly Tensor Codecs*. The 58th IEEE/ACM International Symposium on Microarchitecture (MICRO), 2025.
Best Paper Award.
4. Hancheng Ye, Zhengqi Gao, Mingyuan Ma, Qinsi Wang, Yuzhe Fu, Ming-Yu Chung, Yueqian Lin, Zhijian Liu, Jianyi Zhang, Danyang Zhuo, Yiran Chen. *Training-free Online KV-cache Communication for Efficient LLM-based Multi-agent System*. The 39th Annual Conference on Neural Information Processing Systems (NeurIPS), 2025.
5. Shuowei Jin, Xueshen Liu, Yongji Wu, Haizhong Zheng, Qingzhao Zhang, Atul Prakash, Matthew Lentz, Danyang Zhuo, Feng Qian, Zhuoqing Mao. *Plato: Plan to Efficient Decode for Large Language Model Inference*. The 2nd Conference on Language Modeling (COLM), 2025.
6. Ying Sheng, Shiyi Cao, Dacheng Li, Banghua Zhu, Zhuohan Li, Danyang Zhuo, Joseph E. Gonzalez, Ion Stoica. *Fairness in Serving Large Language Models*. The 18th USENIX Symposium on Operating Systems Design and Implementation (OSDI), 2024.
7. Yongji Wu, Yechen Xu, Jingrong Chen, Zhaodong Wang, Ying Zhang, Matthew Lentz, Danyang Zhuo. *MCCS: A Service-based Approach to Collective Communication for Multi-Tenant Cloud*. In Proceedings of the Conference of the ACM Special Interest Group on Data Communication (SIGCOMM), 2024.

8. Jiaqi Lou, Xinhao Kong, Jinghang Huang, Wei Bai, Nam Sung Kim, Danyang Zhuo. *Hardware-Assisted RDMA Performance Isolation for Public Clouds*. The 21th USENIX Symposium on Networked Systems Design and Implementation (NSDI), 2024.
9. Jinghan Huang, Jiaqi Lou, Srikar Vanavasam, Xinhao Kong, Houxiang Ji, Ipoom Jeong, Eun Kyung Lee, Danyang Zhuo, Nam Sung Kim. *HAL: Hardware-assisted Load Balancing for Energy-efficient SNIC-Host Cooperative Computing*. The 51th IEEE/ACM International Symposium on Computer Architecture (ISCA), 2024.
10. Samantha Miller, Anirudh Kumar, Tanay Vakharia, Ang Chen, Danyang Zhuo, Thomas E. Anderson. *Enoki: High Velocity Linux Kernel Scheduler Development*. The 19th European Conference on Computer Systems (EuroSys), 2024.
11. Lequn Chen, Zihao Ye, Yongji Wu, Danyang Zhuo, Luis Ceze, Arvind Krishnamurthy. *Punica: Multi-Tenant LoRA Serving*. The 7th Conference on Machine Learning and Systems (MLSys), 2024.
12. Jingrong Chen, Yongji Wu, Shihan Lin, Yechen Xu, Xinhao Kong, Thomas E. Anderson, Matthew Lentz, Xiaowei Yang, Danyang Zhuo. *Remote Procedure Call as a Managed System Service*. The 20th USENIX Symposium on Networked Systems Design and Implementation (NSDI), 2023.
13. Xinhao Kong, Jingrong Chen, Wei Bai, Yechen Xu, Mahmoud Elhaddad, Shachar Raindel, Jitendra Padhye, Alvin R. Lebeck, Danyang Zhuo. *Understanding RDMA Microarchitecture Resources for Performance Isolation*. The 20th USENIX Symposium on Networked Systems Design and Implementation (NSDI), 2023.
14. Yongji Wu, Matthew Lentz, Danyang Zhuo, Yao Lu. *Serving and Optimizing Machine Learning Workflows on Heterogeneous Infrastructures*. The 49th International Conference on Very Large Data Bases (VLDB), 2023.
15. Lianke Qin, Rajesh Jayaram, Elaine Shi, Zhao Song, Danyang Zhuo, Shumo Chu. *Adore: Differentially Oblivious Relational Database Operators*. The 49th International Conference on Very Large Data Bases (VLDB), 2023.
16. Hongyi Liu, Jiarong Xing, Yibo Huang, Danyang Zhuo, Srinivas Devadas, Ang Chen. *Remote Direct Memory Introspection*. The 32nd USENIX Security Symposium (USENIX Security), 2023. **Distinguished Paper Award**.
17. Josh Alman, Jiehao Liang, Zhao Song, Ruizhe Zhang, Danyang Zhuo. *Bypass Exponential Time Preprocessing: Fast Neural Network Training via Weight-Data Correlation Preprocessing*. The 37th Conference on Neural Information Processing Systems (NeurIPS), 2023.
18. Xiangfeng Zhu, Guozhen She, Bowen Xue, Yu Zhang, Yongsu Zhang, Xuan Kelvin Zou, Xiongchun Duan, Peng He, Arvind Krishnamurthy, Matthew Lentz, Danyang Zhuo, Ratul Mahajan. *Dissecting Overheads of Service Mesh Sidecars*. The 14th ACM Symposium on Cloud Computing (SoCC), 2023.
19. Lianke Qin, Zhao Song, Lichen Zhang, Danyang Zhuo. *An Online and Unified Algorithm for Projection Matrix Vector Multiplication with Application to Empirical Risk Minimization*. The 26th International Conference on Artificial Intelligence and Statistics (AISTATS), 2023.
20. Lianmin Zheng, Zhuohan Li, Hao Zhang, Yonghao Zhuang, Zhifeng Chen, Yanping Huang, Yida Wang, Yuanzhong Xu, Danyang Zhuo, Eric P. Xing, Joseph E. Gonzalez, Ion Stoica. *Alpa: Automating Inter- and Intra-Operator Parallelism for Distributed Deep Learning*. The 16th USENIX Symposium on Operating Systems Design and Implementation (OSDI), 2022.

21. Xinhao Kong, Yibo Zhu, Huaping Zhou, Zuo Jiang, Jianxi Ye, Chuanxiong Guo, Danyang Zhuo. *Collie: Finding Performance Anomalies in RDMA Subsystems*. The 19th USENIX Symposium on Networked Systems Design and Implementation (NSDI), 2022.
22. Jingrong Chen, Hong Zhang, Wei Zhang, Liang Luo, Jeffrey Chase, Ion Stoica, Danyang Zhuo. *NetHint: White-Box Networking for Multi-Tenant Data Centers*. The 19th USENIX Symposium on Networked Systems Design and Implementation (NSDI), 2022.
23. Danyang Zhuo, Kaiyuan Zhang, Zuhuan Li, Siyuan Zhuang, Stephanie Wang, Ang Chen, Ion Stoica. *Rearchitecting In-Memory Object Stores for Low Latency*. The 48th International Conference on Very Large Data Bases (VLDB), 2022.
24. Shunhua Jiang, Yunze Man, Zhao Song, Zheng Yu, Danyang Zhuo. *Fast Graph Neural Tangent Kernel via Kronecker Sketching*. The 36th AAAI Conference on Artificial Intelligence (AAAI), 2022.
25. Siyuan Zhuang, Zuhuan Li, Danyang Zhuo, Stephanie Wang, Eric Liang, Robert Nishihara, Philipp Moritz, Ion Stoica. *Hoplite: Efficient and Fault-Tolerant Collective Communication for Task-Based Distributed Systems*. In Proceedings of the Conference of the ACM Special Interest Group on Data Communication (SIGCOMM), 2021.
26. Samantha Miller, Kaiyuan Zhang, Mengqi Chen, Ryan Jennings, Ang Chen, Danyang Zhuo, Thomas E. Anderson. *High Velocity Kernel File Systems with Bento*. The 19th USENIX Conference on File and Storage Technologies (FAST), 2021.
Best Paper Award.
27. Zuhuan Li, Siyuan Zhuang, Shiyuan Guo, Danyang Zhuo, Hao Zhang, Dawn Song, Ion Stoica. *TeraPipe: Token-Level Pipeline Parallelism for Training Large-Scale Language Models*. The 38th International Conference on Machine Learning (ICML), 2021.
28. Sitan Chen, Xiaoxiao Li, Zhao Song, Danyang Zhuo. *On InstaHide, Phase Retrieval, and Sparse Matrix Factorization*. The 9th International Conference on Learning Representations (ICLR), 2021.
29. Shumo Chu, Danyang Zhuo, Elaine Shi, T-H. Hubert Chan. *Differentially Oblivious Database Joins: Overcoming the Worst-Case Curse of Fully Oblivious Algorithms*. The 2nd Information-Theoretic Cryptography conference (ITC), 2021.
30. Lianmin Zheng, Chengfan Jia, Minmin Sun, Zhao Wu, Cody Hao Yu, Ameer Haj-Ali, Yida Wang, Jun Yang, Danyang Zhuo, Koushik Sen, Joseph E. Gonzalez, Ion Stoica. *Anso: Generating High-Performance Tensor Programs for Deep Learning*. The 14th USENIX Symposium on Operating Systems Design and Implementation (OSDI), 2020.
31. Kaiyuan Zhang, Danyang Zhuo, Arvind Krishnamurthy. *Gallium: Automated Software Middlebox Offloading to Programmable Switches*. In Proceedings of the Conference of the ACM Special Interest Group on Data Communication (SIGCOMM), 2020.
32. Kaiyuan Zhang, Danyang Zhuo, Aditya Akella, Arvind Krishnamurthy, Xi Wang. *Automated Verification of Customizable Middlebox Properties with Gravel*. The 17th USENIX Symposium on Networked Systems Design and Implementation (NSDI), 2020.
33. Danyang Zhuo, Kaiyuan Zhang, Yibo Zhu, Hongqiang Harry Liu, Matthew Rockett, Arvind Krishnamurthy, Thomas E. Anderson. *Slim: OS Kernel Support for a Low-Overhead Container Overlay Network*. The 16th USENIX Symposium on Networked Systems Design and Implementation (NSDI), 2019.

34. Danyang Zhuo, Monia Ghobadi, Ratul Mahajan, Klaus-Tycho Förster, Arvind Krishnamurthy and Thomas E. Anderson. *Understanding and Mitigating Packet Corruption in Data Center Networks*. In Proceedings of the Conference of the ACM Special Interest Group on Data Communication (SIGCOMM), 2017.
35. Danyang Zhuo, Monia Ghobadi, Ratul Mahajan, Amar Phanishayee, Xuan Kelvin Zou, Hang Guan, Arvind Krishnamurthy and Thomas E. Anderson. *RAIL: A Case for Redundant Arrays of Inexpensive Links in Data Center Networks*. The 14th USENIX Symposium on Networked Systems Design and Implementation (NSDI), 2017.
36. Vincent Liu, Danyang Zhuo, Simon Peter, Arvind Krishnamurthy and Thomas E. Anderson. *Subways: A Case for Redundant, Inexpensive Data Center Edge Links*. The 13th International Conference on emerging Networking EXperiments and Technologies (CoNEXT), 2015.

Workshop Papers

1. Jianxing Qin, Alexander Du, Danfeng Zhang, Matthew Lentz, Danyang Zhuo. *Can Large Language Models Verify System Software? A Case Study Using FSCQ as a Benchmark*. The 20th Workshop on Hot Topics in Operating Systems (HotOS), 2025.
2. Xiangfeng Zhu, Yang Zhou, Yuyao Wang, Xiangyu Gao, Arvind Krishnamurthy, Sam Kumar, Ratul Mahajan, Danyang Zhuo. *Rethinking RPC Communication for Microservices-based Applications*. The 20th Workshop on Hot Topics in Operating Systems (HotOS), 2025.
3. Xinhao Kong, Jiaqi Lou, Wei Bai, Nam Sung Kim, Danyang Zhuo. *Towards a Manageable Intra-Host Network*. The 19th Workshop on Hot Topics in Operating Systems (HotOS), 2023.
4. Xiangfeng Zhu, Weixin Deng, Banruo Liu, Jingrong Chen, Yongji Wu, Thomas E. Anderson, Arvind Krishnamurthy, Ratul Mahajan, Danyang Zhuo. *Application Defined Networks*. The 22nd ACM Workshop on Hot Topics in Networks (HotNets), 2023.
5. Jialin Li, Samantha Miller, Danyang Zhuo, Ang Chen, Jon Howell, Thomas E. Anderson. *An Incremental Path Towards a Safe OS Kernel*. The 18th Workshop on Hot Topics in Operating Systems (HotOS), 2021.
6. John Snyder, Alvin R. Lebeck, Danyang Zhuo. *RDMA Congestion Control: It's Only for the Compliant*. Cloud @ MICRO, 2021.
7. Samantha Miller, Kaiyuan Zhang, Danyang Zhuo, Shabin Xu, Arvind Krishnamurthy, Thomas E. Anderson. *Practical Safe Linux Kernel Extensibility*. The 17th Workshop on Hot Topics in Operating Systems (HotOS), 2019.
8. Danyang Zhuo, Qiao Zhang, Xin Yang, Vincent Liu. *Canaries in the Network*. The 15th ACM Workshop on Hot Topics in Networks (HotNets), 2016.
9. Danyang Zhuo, Qiao Zhang, Vincent Liu, Arvind Krishnamurthy, Thomas E. Anderson. *Rack-level Congestion Control*. The 15th ACM Workshop on Hot Topics in Networks (HotNets), 2016.
10. Danyang Zhuo, Qiao Zhang, Dan Ports, Arvind Krishnamurthy, Thomas E. Anderson. *Machine Fault Tolerance for Reliable Datacenter Systems*. The 5th Asia-Pacific Workshop on Systems (APSys), 2014.

Journal Papers

1. John Snyder, Alvin R. Lebeck, Danyang Zhuo. *RDMA Congestion Control: It's Only for the Compliant*. IEEE Micro, 2022.

Invited Papers

1. Samantha Miller, Kaiyuan Zhang, Mengqi Chen, Ryan Jennings, Ang Chen, Danyang Zhuo, Thomas E. Anderson. *High Velocity Kernel File Systems with Bento*. USENIX ;login:, 2021.

Patents

1. Monia Ghobadi, Ratul Mahajan, Amar Phanishayee, Danyang Zhuo, Xuan Kelvin Zou. *Data Center Topology Having Multiple Classes of Reliability*. US Patent 20170302565A1. WIPO Patent 2017180450A1.

Mentoring

Current PhD Students

- Zhuo Chen
 - Duke CS Research Initiation Project Award (2024)
- Jianxing Qin
- Yechen Xu
- Yuncheng Yao (with Tingjun Chen)

Past PhD Students

- Jingrong Chen (2025), Duke University
 - Dissertation: *Flexible and Efficient Application-Networking Co-Design in Cloud Datacenters*
 - Current Appointment: Uber
 - Duke CS Research Initiation Project Award (2022)
 - Duke CS Teaching Assistant Award (2022)
- Yongji Wu (with Matthew Lentz, 2024), Duke University
 - Dissertation: *Optimizing Distributed Workloads with Infrastructure-managed Communication and Deployment*
 - Current Appointment: Postdoc at UC Berkeley
 - IEEE/ACM MICRO Best Paper Award (2025)
- Xinhao Kong (2024), Duke University
 - Dissertation: *Towards Large-Scale RDMA Networking Without Performance Anomalies*
 - Current Appointment: NVIDIA
 - Duke CS Outstanding Ph.D. Preliminary Research Exam (2024)
 - Duke CS Research Initiation Project Award (2023)

- Duke CS Teaching Assistant Award (2023)
- Samantha Miller (with Tom Anderson, 2023), University of Washington
 - Dissertation: *High Velocity Operating Systems Development*
 - Current Appointment: Databricks
 - USENIX FAST Best Paper Award (2021)

Past Master Students

- Guozhen She (2022), Duke University
 - Project: *Understanding the Design and Implementation of Service Meshes.*
 - First Appointment: Amazon
- Wei Zhang (2022), Duke University
 - Project: *Does Single-Node Optimization Help Distributed In-Memory Object Store?*
 - First Appointment: Microsoft
- Zhangzhang Yue (2022), Duke University
 - Project: *Balancing Bandwidth and Accuracy in Distributed Video Analytics Systems.*
 - Duke CS Master Project/Thesis Award (2022)
 - First Appointment: SmartNews

Ph.D. Dissertation Committee

- Shihan Lin (2025), Duke University
 - Dissertation: *Securing Web Content Distribution in the Presence of Third-Party Content Delivery Networks.*
- Yitu Wang (2025), Duke University
 - Dissertation: *Near Data Processing for Data-Intensive Machine Learning Applications.*
- Sifei Luan (2024), University of California - Berkeley
 - Dissertation: *An Extensive Architecture for Distributed Heterogeneous Processing.*
- Ceyu Xu (2024), Duke University
 - Dissertation: *Artificial Intelligence for Intelligent Computer Architecture.*
- Siyuan Zhuang (2024), University of California - Berkeley
 - Dissertation: *Providing Efficient Fault Tolerance in Distributed Systems.*
- Shiyu Li (2024), Duke University
 - Dissertation: *Joint Optimization of Algorithms, Hardware, and Systems for Efficient Deep Neural Network.*
- Lequn Chen (2024), University of Washington

- Dissertation: *Multi-Tenant Machine Learning Model Serving Systems on GPU Clusters*.
- Xiao Zhang (2023), Duke University
 - Dissertation: *Proactive and Passive Performance Optimization of IP Anycast*.
- Jack Snyder (2022), Duke University
 - Dissertation: *Improving Congestion Control Convergence in RDMA Networks*.
- Kaiyuan Zhang (2021), University of Washington
 - Dissertation: *Automated Analysis of Correct and Efficient Execution of Software Middleboxes*.

Invited Talks

- **Phantora: Maximizing Code Reuse in Simulation-based Machine Learning System Performance Estimation**
 - Google Networking Research Summit Oct 2025
- **Towards Efficient and Manageable Host Networking.**
 - Rice University Nov 2023
 - Georgetown University Nov 2023
 - University of Maryland Nov 2023
- **Remote Procedure Call as a Managed System Service.**
 - University of Minnesota Oct 2023
 - UW FOCI Workshop on Application Networking May 2023
- **Systematic Testing of High-Speed RDMA Networks.**
 - Meta Data Application for Better Infrastructure Conference Dec 2022
 - Cornell University Oct 2022
 - Microsoft Research Sep 2022
 - Meta Infrastructure Data Science Faculty Workshop Aug 2022
- **In-Memory Object Stores for Low Latency.**
 - VLDB Sep 2022
- **Collie: Finding Performance Anomalies in RDMA Subsystems.**
 - Google Networking Research Summit Mar 2022
 - Microsoft Azure Sep 2021
- **Towards Efficient Cloud Systems for Data-Intensive Applications.**
 - Rice University Jun 2021
 - Duke CS+ Undergraduate Summer Research Program Jun 2021
 - IBM Feb 2021

- **Towards Efficient and Reliable Data Center Systems.**
 - Yale University Apr 2019
 - Purdue University Apr 2019
 - University of Virginia Mar 2019
 - Duke University Mar 2019
 - Rutgers University Mar 2019
 - Microsoft Research Mar 2019
 - Pennsylvania State University Feb 2019
 - University of Minnesota Feb 2019
- **Slim: OS Kernel Support for a Low-Overhead Container Overlay Network.**
 - Princeton University Jun 2020
 - University of California - Berkeley Nov 2019
 - USENIX NSDI Feb 2019
- **Understanding and Mitigating Packet Corruption in Data Center Networks.**
 - ACM SIGCOMM Aug 2017
- **RAIL: A Case for Redundant Arrays of Inexpensive Links in Data Center Networks.**
 - USENIX NSDI Mar 2017

Teaching

- **Spring 2026: Systems for Machine Learning (CompSci 590.06)**
- **Fall 2025: Introduction to Operating Systems (CompSci 310)**
- **Spring 2025: Systems for Machine Learning (CompSci 590.05)**
 - Instructor Evaluation: 4.70/5
- **Spring 2024: Introduction to Computer Systems (CompSci 210)**
 - Instructor Evaluation: 1.95/5
- **Spring 2023: Distributed Systems (CompSci 512)**
 - Instructor Evaluation: 4.27/5
- **Fall 2022: Introduction to Operating Systems (CompSci 310)**
 - Instructor Evaluation: 3.83/5
- **Spring 2022: Data Center Systems (CompSci 590.04)**
 - Instructor Evaluation: 4.00/5
- **Fall 2021: Introduction to Operating Systems (CompSci 310)**
 - Instructor Evaluation: 3.69/5

- Fall 2020: Advanced Operating Systems (CompSci 510)

- Instructor Evaluation: 4.60/5

Service

Organizer

- 2022: Co-Chair of SIGCOMM Artifact Evaluation Committee

Technical Program Committee

- 2026: ASPLOS, MLSys, NSDI, SIGMOD
- 2025: APNET, NSDI, SIGCOMM, SIGMOD
- 2024: APNET, CoNEXT, NSDI, SIGMOD
- 2023: APNET, NSDI, SIGCOMM
- 2022: APNET, FAST, NSDI, SIGCOMM
- 2021: CoNEXT
- 2020: SIGCOMM

Proposal Review Panel

- 2025: NSF NeTS
- 2022: NSF NeTS

Department Service

- Duke CS Colloquium Committee Chair (2025-now)
- Duke CS PhD Admission Committee Member (2022-2025)
- Duke CS Communication Committee Member (2020-2022)

University Service

- Duke Information Technology Advisory Council Member (2023-2024)