Yang Zhou

yangzhou1997.github.io
yangzhou.rpc@gmail.com ⋄ +1 617 599 8532

Research Interests

Machine learning systems, networked systems

Education

Harvard University, Cambridge, MA, USA

Ph.D. in Computer Science — *June 2024*
M.S. in Computer Science — *November 2021*
Thesis: Network-Application Co-design for Efficient Datacenters
Advisors: Minlan Yu and James Mickens

Peking University, Beijing, China

B.S. in Computer Science — *July 2018*
Thesis: Towards Faster and More Accurate Data Stream Processing
Advisor: Tong Yang

Employment

University of California, Davis, Assistant Professor of Computer Science — *July 2025 -*

University of California, Berkeley, Postdoctoral researcher in Sky Computing Lab — *July 2024 - now*

Project: UCCL: an Efficient Collective Communication Library for GPUs
Supervisor: Ion Stoica

Google SRG and NetInfra, Research Intern — *June 2021 - May 2023*

VMware Research, Research Intern — *July 2020 - September 2020*

Meta/Facebook, Research Collaborator — *November 2019 - May 2020*

SenseTime, Software Engineering Intern — *March 2018 - May 2018*

Professional Service

Leadership:

- Co-Chair: ACM SIGCOMM Artifact Evaluation 2024

Program Committees:

- USENIX NSDI: 2026
- USENIX OSDI: 2025
- ACM ASPLOS: 2026
- ACM SIGCOMM Workshop on eBPF and Kernel Extensions 2024, 2025
- ACM SIGCOMM Workshop on Networks for AI Computing 2025
- ACM SIGCOMM Poster/Demo 2023, 2024, 2025
- IEEE INFOCOM: Workshop on Networking Algorithms 2020

PUBLICATIONS

Conference

[1] Xiangfeng Zhu, **Yang Zhou**, Yuyao Wang, Xiangyu Gao, Arvind Krishnamurthy, Sam Kumar, Ratul Mahajan, Danyang Zhuo. [Link]
Rethinking RPC Communication for Microservices-based Applications.
HotOS 2025.

[2] Xuanlin Jiang, **Yang Zhou**, Shiyi Cao, Ion Stoica, Minlan Yu.
NEO: Saving GPU Memory Crisis with CPU Offloading for Online LLM Inference. [link]
MLSys 2025.

[3] Zhongjie Chen, Qingkai Meng, ChonLam Lao, Yifan Liu, Fengyuan Ren, Minlan Yu, **Yang Zhou**.
eTran: Extensible Kernel Transport with eBPF. [link]
USENIX NSDI 2025.

[4] **Yang Zhou**, Mark Wilkening, James Mickens, and Minlan Yu.
SmartNIC Security Isolation in the Cloud with S-NIC. [link]
ACM EuroSys 2024.

[5] **Yang Zhou**, Xingyu Xiang, Matthew Kiley, Sowmya Dharanipragada, and Minlan Yu.
DINT: Fast In-Kernel Distributed Transactions with eBPF. [link]
USENIX NSDI 2024.

[6] **Yang Zhou**, Zezhou Wang, Sowmya Dharanipragada, and Minlan Yu.
Electrode: Accelerating Distributed Protocols with eBPF. [link]
USENIX NSDI 2023.

[7] **Yang Zhou**, Hassan Wassel, Sihang Liu, Jiaqi Gao, James Mickens, Minlan Yu, Chris Kennelly, Paul Turner, David Culler, Hank Levy, and Amin Vahdat.
Carbink: Fault-Tolerant Far Memory. [link]
USENIX OSDI 2022.

[8] **Yang Zhou**, Ying Zhang, Minlan Yu, Guangyu Wang, Dexter Cao, Eric Sung, and Starsky Wong.
Evolvable Network Telemetry at Facebook. [link]
USENIX NSDI 2022.

[9] **Yang Zhou**, Tong Yang, Jie Jiang, Bin Cui, Minlan Yu, Xiaoming Li, and Steve Uhlig.
Cold Filter: A Meta-Framework for Faster and More Accurate Stream. Processing [link]
ACM SIGMOD 2018.

[10] Tong Yang, Jie Jiang, Peng Liu, Qun Huang, Junzhi Gong, **Yang Zhou**, Rui Miao, Xiaoming Li, and Steve Uhlig.
Elastic Sketch: Adaptive and Fast Network-Wide Measurements. [link]
ACM SIGCOMM 2018.

[11] Omid Alipourfard, Masoud Moshref, **Yang Zhou**, Tong Yang, and Minlan Yu.
A Comparison of Performance and Accuracy of Measurement Algorithms in Software. [link]
ACM Symposium on SDN Research (SOSR) 2018.

[12] Xiangyang Gou, Chenxingyu Zhao, Tong Yang, Lei Zou, **Yang Zhou**, Yibo Yan, Xiaoming Li, and Bin Cui.
Single Hash: Use One Hash Function to Build Faster Hash Based Data Structures. [link]
IEEE International Conference on Big Data and Smart Computing (BigComp) 2018.

[13] Tong Yang, **Yang Zhou**, Hao Jin, Shigang Chen, and Xiaoming Li.

Pyramid Sketch: A Sketch Framework for Frequency Estimation of Data Streams. [link]
VLDB 2017.

[14] **Yang Zhou**, Peng Liu, Hao Jin, Tong Yang, Shoujiang Dang, and Xiaoming Li.
One Memory Access Sketch: A More Accurate and Faster Sketch for Per-Flow Measurement. [link]
IEEE Global Communications Conference (Globecom) 2017.

[15] Junzhi Gong, Tong Yang, **Yang Zhou**, Dongsheng Yang, Shigang Chen, Bin Cui, and Xiaoming Li.
ABC: A Practicable Sketch Framework for Non-Uniform Multisets. [link]
IEEE International Conference on Big Data (BigData) 2017.

Workshop and Demo

[16] **Yang Zhou**, Hao Jin, Peng Liu, Haowei Zhang, Tong Yang, and Xiaoming Li.
Accurate Per-Flow Measurement with Bloom Sketch. [link]
IEEE International Conference on Computer Communications Workshops (INFOCOM WKSHPS) 2018.

Journal

[17] Zhuochen Fan, Gang Wen, Zhipeng Huang, **Yang Zhou**, Qiaobin Fu, Tong Yang, Alex X Liu, and Bin Cui.
On the Evolutionary of Bloom Filter False Positives - An Information Theoretical Approach to Optimizing Bloom Filter Parameters. [link]
IEEE Transactions on Knowledge & Data Engineering 2022.

[18] Yuanpeng Li, Xiang Yu, Yilong Yang, **Yang Zhou**, Tong Yang, Zhuo Ma, and Shigang Chen.
Pyramid Family: Generic Frameworks for Accurate and Fast Flow Size Measurement. [link]
IEEE/ACM Transactions on Networking 2021.

[19] Tong Yang, Jie Jiang, **Yang Zhou**, Long He, Jinyang Li, Bin Cui, Steve Uhlig, and Xiaoming Li.
Fast and Accurate Stream Processing by Filtering the Cold. [link]
The VLDB Journal 2019.

[20] Tong Yang, Jie Jiang, Peng Liu, Qun Huang, Junzhi Gong, **Yang Zhou**, Rui Miao, Xiaoming Li, and Steve Uhlig.
Adaptive Measurements Using One Elastic Sketch. [link]
IEEE/ACM Transactions on Networking 2019.

[21] **Yang Zhou**, Omid Alipourfard, Minlan Yu, and Tong Yang.
Accelerating Network Measurement in Software. [link]
ACM SIGCOMM Computer Communication Review 2018.

Preprints

[22] Jiarong Xing, Yifan Qiao, Simon Mo, Xingqi Cui, Gur-Eyal Sela, **Yang Zhou**, Joseph Gonzalez, Ion Stoica.
Towards Efficient and Practical GPU Multitasking in the Era of LLM. [link]
Arxiv Aug 2025

[23] Yichuan Wang, Shu Liu, Zhifei Li, Yongji Wu, Ziming Mao, Yilong Zhao, Xiao Yan, Zhiying Xu, **Yang Zhou**, Ion Stoica, Sewon Min, Matei Zaharia, Joseph E. Gonzalez.
LEANN: a Low-Storage Vector Index. [link]
Arxiv June 2025

[24] **Yang Zhou**, Zhongjie Chen, Ziming Mao, ChonLam Lao, Shuo Yang, Pravein Govindan Kannan, Jiaqi Gao, Yilong Zhao, Yongji Wu, Kaichao You, Fengyuan Ren, Zhiying Xu, Costin Raiciu, Ion Stoica.
UCCL: an Efficient Collective Communication Library for GPUs. [link]
Arxiv April 2025

[25] Shiyi Cao, Yichuan Wang, Ziming Mao, Pin-Lun Hsu, Liangsheng Yin, Tian Xia, Dacheng Li, Shu Liu, Yineng Zhang, **Yang Zhou**, Ying Sheng, Joseph Gonzalez, Ion Stoica.
Locality-Aware Fair Scheduling in LLM Serving. [link]
Arxiv Jan 2025

[26] Yilong Zhao, Shuo Yang, Kan Zhu, Lianmin Zheng, Baris Kasikci, Yifan Qiao, **Yang Zhou**, Jiarong Xing, Ion Stoica.
BlendServe: Optimizing Offline Inference for Auto-regressive Large Models with Resource-aware Batching. [link]
Arxiv Nov 2024

[27] Yifan Qiao, Shu Anzai, Shan Yu, Haoran Ma, Shuo Yang, Yang Wang, Miryung Kim, Yongji Wu, **Yang Zhou**, Jiarong Xing, Joseph Gonzalez, Ion Stoica, Harry Xu.
ConServe: Harvesting GPUs for Low-Latency and High-Throughput Large Language Model Serving. [link]
Arxiv Oct 2024

[28] Shuo Yang, Ying Sheng, Yilong Zhao, Joseph Gonzalez, **Yang Zhou**, Ion Stoica, Lianmin Zheng.
Post-Training Sparse Attention with Double Sparsity. [link]
Arxiv Aug 2024

Open Source Software

- UCCL, ultra and unified CCL for GPU communication, 490+ stars
  https://github.com/uccl-project/uccl
- LEANN, the smallest vector index in the world for RAG, 1100+ stars
  https://github.com/yichuan-w/LEANN
- NEO, an LLM inference engine built to save the GPU memory by CPU offloading
  https://github.com/NEO-MLSys25/NEO
- eTran, implementing reliable network transports in the Linux kernel with eBPF
  https://github.com/eTran-NSDI25/eTran
- DINT, running distributed transactions in the Linux kernel with eBPF
  https://github.com/DINT-NSDI24/DINT
- Electrode, running Paxos consensus in the Linux kernel with eBPF
  https://github.com/Electrode-NSDI23/Electrode

Students

Current Students

- Shuang Ma (PhD) *2025 - now*
- Yihan Zhang (PhD) *2025 - now*

Mentored Students

- Zhongjie Chen, Tsinghua University PhD *2024 - now*
  Extensible kernel transport (NSDI 2025, [3]).
- Xuanlin Jiang, Peking University undergraduate → Harvard PhD *2024*
  CPU offloading for online LLM inference (MLSys 2025, [2]).
- Matt Kiley, Harvard College undergraduate → Clockwork Systems *2023*
  Accelerating distributed transactions using eBPF (NSDI 2024, [5]).

- Yunxi Shen, Tsinghua University undergraduate → Cornell PhD *2023*
  Resource-efficient job scheduling in data centers.
- Xingyu Xiang, Peking University undergraduate → Harvard PhD *2023*
  Accelerating distributed transactions using eBPF (NSDI 2024, [5]).
- Zezhou Wang, Peking University undergraduate → University of Washington PhD *2022*
  Accelerating Paxos using eBPF (NSDI 2023, [6]).

TALKS

- UCCL: an Efficient Collective Communication Library for GPUs
  Meta, ByteDance (Seed), SJTU IPADS *July 2025*
  ByteDance (Networking), NVIDIA *June 2025*
  UC Berkeley SkyLab Summer Retreat, Broadcom *May 2025*
  UC Berkeley SkyLab Winter Retreat *January 2025*
- Network-Application Co-design for Efficient Datacenters
  University of Toronto *April 2024*
  NYU, Brown, UC Irvine, UWaterloo, UC Davis, Boston University *March 2024*
  UC Santa Cruz, University of Virginia, Purdue *February 2024*
- Electrode: Accelerating Distributed Protocols with eBPF
  Duke University, ACE Center for Evolvable Computing, Google, USENIX NSDI *April 2023*
  Columbia University *March 2023*
- Carbink: Fault-Tolerant Far Memory
  Cornell University *November 2023*
  WORDS workshop *November 2022*
  Microsoft Research Redmond, USENIX OSDI *July 2022*
  Google *March & June 2022*
- Evolvable Network Telemetry at Facebook
  USENIX NSDI *April 2022*
  Boston University, Meta *March 2022*
- Cold Filter: A Meta-Framework for Faster and More Accurate Stream Processing
  Harvard University *October 2018*

TEACHING EXPERIENCE

- **Guest Lecture** on far memory, CS294-252: Architectures and Systems for Warehouse-Scale Computers, UC Berkeley *Nov 2023*
- **Teaching Assistant** for Prof. Minlan Yu, CS145: Networking at Scale, Harvard University *Spring 2021*
- **Teaching Assistant** for Prof. Tong Yang, Algorithm Design and Analysis, Peking University *Fall 2018*

PATENTS

- **Yang Zhou**, Hassan Wassel, Minlan Yu, Hank Levy, David Culler, and Amin Vahdat. "Fault Tolerant Disaggregated Memory". Pending (US20230185666A1), December 2022.

AWARDS AND HONORS

- Google Ph.D. Fellowship in Systems and Networking *2022*

- Finalist, Meta Ph.D. Fellowship in Networking *2022*
- Graduate Fellowship, Harvard University *2018*
- Excellent Bachelor Thesis (10/327), School of EECS, Peking University *2018*
- New Academic Star Award (1/193), School of EECS, Peking University *2018*
- Arawana Scholarship (2/193), Peking University *2017*
- Pinyou Hudong Scholarship, School of EECS, Peking University *2016*
- May Fourth Scholarship, Peking University *2015*