Yongzhou Chen

✆ +1 217 693 1178| ✉ yc28@illinois.edu | Homepage: https://yongzhouc.com/

EDUCATION

University of Illinois, Urbana-Champaign — IL, USA
Ph.D. in Electrical and Computer Engineering — Sep 2019 – May 2025
- GPA **3.90** / 4.0, Advisor: Prof. Radhika Mittal

University of Science and Technology of China — Hefei, China
B.S. in Computer Science with honors — Sep 2015 – Jun 2019
- Top **5**% with *summa cum laude*

PUBLICATIONS

- **Harp: Handover Assisted 5G RAN Slicing across Multiple gNBs** — In submission
  Yongzhou Chen, Taimoor Tariq, Haitham Hassanieh, Radhika Mittal
- **SiRAN: Channel-Aware 5G RAN Slicing Across Multiple Interfering Cells** — In submission
  Taimoor Tariq, **Yongzhou Chen**, Haitham Hassanieh, Radhika Mittal
- **Octopus: In-Network Content Adaptation to Control Congestion on 5G Links** — SEC 2023
  Yongzhou Chen, Ammar Tahir, Francis Y. Yan, Radhika Mittal
- **Channel-Aware 5G RAN Slicing with Customizable Schedulers** — NSDI 2023
  Yongzhou Chen, Ruihao Yao, Haitham Hassanieh, Radhika Mittal
- **FedSS: Federated Learning with Smart Selection of Clients** — FLSys 2023(with MLSys)
  Yongzhou Chen[*], Ammar Tahir[*], Prashanti Nilayam

WORK EXPERIENCE

Meta Platforms — CA, USA
Research Scientist, AI System SW/HW Co-Design — March 2025 – Present
- Research & Development in the AI and Systems Co-Design team
- Work on collective communications for various hardware systems(Nvidia and AMD GPUs) to support GenAI and Ads workloads

University of Illinois Urbana-Champaign — IL, USA
Research Assistant — Sep 2019 – Jan 2025
- Conducted research in 5G networks and edge computing with Prof. Radhika Mittal
- Made contributions in load-balancing, resource allocation and congestion control in 5G networks to improve its efficiency, fairness and QoS-awareness
- Published papers in top peer-reviewed network conferences such as NSDI, SEC, and MLSys

Microsoft Research — WA, USA
Research Intern — May 2023 – Aug 2023
Hosts: Dr. Ilias Marinos, Dr. Prateesh Goyal
- Designed and implemented Torrent, a multipath datacenter protocol, aimed at optimizing high-throughput and load-balancing requirements for large-scale machine learning training workloads
- Torrent is sender-driven, compatible with commodity switches, and preserves RDMA semantics
- Evaluation in the htsim simulator shows that our protocol reduces the completion time of all-to-all collective operations by 12% compared with state-of-the-art single-path protocols

Google Cloud — CA, USA
PhD Software Engineer Intern — May 2022 – Aug 2022
Hosts: Abhiram Ravi, Tyler Griggs, Dr. Masoud Moshref
- Implemented an advanced congestion control algorithm based on Swift which leverages the in-network telemetry to ramp up fast when there's available bandwidth in the fabric
- Extensively evaluated the designed congestion control algorithm in RPC benchmarks and storage benchmarks, and shows 15% operation transfer latency reduction with bursty background traffic
- Designed and implemented an informed PLB(protective load balancing) system that collects fabric available bandwidth information and reroutes flows effectively

RESEARCH EXPERIENCE

• Network Support for AI/ML Workloads 2024-Present

UIUC

The distributed AI/ML workloads are diverse, ranging from decentralized federated learning to large-scale LLM training. For federated learning, we propose a smart client selection algorithm(FedSS) to reduce the overall training time while preserving data heterogeneity. The results have been published in **FLSys@MLSys'23**. We also explored how to alleviate network congestion for collective communications, which takes significant time in distributed ML training. We propose to dynamically tune the collective algorithm to control how data flows between nodes to avoid congested links. This ongoing work is in submission.

• Resource Allocations and Virtualization in 5G Networks 2022 - 2024

UIUC

5G slicing(or 5G virtualization) refers to allocate the wireless resources to multiple slices owned by different entities(e.g. MVNOs, enterprises, campuses). The slices enter the SLAs with the network operator to define the type of service wanted(e.g. quota of resources, scheduling policies, user priorities, etc). I have proposed multiple innovative algorithms and build 5G slicing systems to: 1. improve the spectral efficiency and link capacity of users(published in **NSDI'23**); 2. balance the load across multiple base stations to improve all slices' performance(submitted to **Mobicom'25**).

• Congestion Control with In-Network Support 2020-2022

UIUC and Google Cloud

Modern congestion controllers in different network scenarios can leverage in-network support to achieve both high throughput and low latency. In data center networks, we proposes to use a practical and simple INT(minimal available bandwidth) to accelerate flow rate convergence. Our protocol built upon Swift further reduces FCT of small RPCs by 15% and is deployed in Google. In 5G networks, I've built a cross-layer system(published in **SEC'23**) that leverages in-network content adaptation to stream real-time applications smoothly over the volatile 5G links. Real-time video applications built upon our system can achieve 2-10× lower tail latency and 17%-50% higher video QoE than WebRTC and Salsify.

PROFESSIONAL SERVICES

- Program Committee, ICPP(54th International Conference on Parallel Processing)
- Reviewer of IEEE Transactions on Networking 2025-Present
- Reviewer of IEEE Network Magazine 2023-Present
- Reviewer of Journal of Network and Computer Applications 2022-Present

AWARDS AND HONORS

- Yi-Min Wang and Pi-Yu Chung Research Award 2024
- Rambus Computer Engineering Fellowship 2023
- Excellent graduate of Anhui Province 2019
- OSDI'18 Student Travel Grant 2018
- Scholarship for Honors Classes in Computer Science 2017-2019
- Outstanding Student Scholarship (Top 10% for academic achievement) 2017-2019
- First Prize in the 5th Student RDMA Programming Competition 2017
- National Scholarship 2016
- First Prize, National Olympiad in Physics in Provinces, China 2014

TEACHING EXPERIENCE

Graduate Student Teaching Assistant UIUC

- *ECE120*: Introduction to Computing Spring 2020

SKILLS AND TECHNIQUES

- *Programming Languages:* C/C++, Python, Java, Go, Rust
- *Skills:* NCCL, Intel DPDK, RDMA Networking, PyTorch, TensorFlow, Kernel Programming, NS-2/3