Yifan Yuan

Email: yifany3@illinois.edu Website: https://yifanyuan3.github.io

Education

- MS/PhD, Department of Electrical and Computer Engineering, UIUC *August 2017 – Present*
  - Major: Computer Architecture and Systems
  - Advisor: Prof. Nam Sung Kim
- Bachelor, Institute of VLSI Design, Zhejiang University *September 2014 – June 2018*
  - Major: Electronic Information Engineering

Research Interests

- Networking hardware and system software for datacenter
- Hardware-software co-design for distributed systems acceleration

Publications

- **Y. Yuan**, O. Alama, J. Fei, J. Nelson, D. R. K. Ports, A. Sapio, M. Canini, N. S. Kim. **Unlocking the Power of Inline Floating-Point Operations on Programmable Switches, *NSDI'22***
- **Y. Yuan**, M. Alian, Y. Wang, R. Wang, I. Kurakin, C. Tai, N. S. Kim. **Don't Forget the I/O When Allocating Your LLC, *ISCA'21 full paper,*** *NSDI'20 poster*
  Code to appear in Intel official RDT (pqos) library
- **Y. Yuan**, Y. Wang, R. Wang, R. Chowdhury, C. Tai, N. S. Kim. **QEI: Query Acceleration Can be Generic and Efficient in the Cloud, *HPCA'21***
- M. Alian, **Y. Yuan**, J. Zhang, R. Wang, M. Jung, N. S. Kim. **Data Direct I/O Characterization for Future I/O System Exploration, *ISPASS'20***
- **Y. Yuan**, Y. Wang, R. Wang, J. Huang. **HALO: Accelerating Flow Classification for Scalable Packet Processing in NFV, *ISCA'19***
- Y. Li, I. Liu, **Y. Yuan**, D. Chen, A. Schwing, J. Huang. **Accelerating Distributed Reinforcement Learning with In-Switch Computing, *ISCA'19 full paper,*** *NSDI'19 poster*
- X. Wang, **Y. Yuan**, Y. Zhou, C. C. Coats, J. Huang. **Project Almanac: A Time-Traveling Solid-State Drive, *EuroSys'19***
- Y. Li, J. Park, M. Alian, **Y. Yuan**, Q. Zheng, P. Pan, R. Wang, A. Schwing, H. Esmaeilzadeh, N. S. Kim. **A Network-Centric Hardware/Algorithm Co-Design to Accelerate Distributed Training of Deep Neural Networks, *MICRO'18***

Patents

- R. Wang, T.-Y. C. Tai, Y. Wang, **Y. Yuan**, S. Paul, M. M. Khellah, S. Gobriel, C. Augustine, M. Ganguli, J.-S. Tsai, E. Verplanke, P. Autee, A. Layek, S. Narayana, B. Ganesh, J. B. Timbadiya, S. K. Muthukumar, R. Iyer, N. Jain, N. D. McDonnell, M. A. Goldschmidt, R. M. Sankaran, N. Ranganathan. **Hardware Assisted Lookup Operations**, *US Patent App. 63/130,663, filed Dec. 2020*
- R. Wang, **Y. Yuan**, Y. Wang, T.-Y. C. Tai, T. Hurson. **Data Consistency and Durability over Distributed Persistent Memory Systems**, *US Patent App. 62/986,094, filed Aug. 2020*
- Y. Wang, R. Wang, T.-Y. C. Tai, **Y. Yuan**, P. Pathak, S. Vedantham, C. Macnamara. **Workload Scheduler for Memory Allocation**, *US Patent App. 16/799,745, filed Feb. 2020*
- R. Wang, A. J. Herdrich, T.-Y. C. Tai, Y. Wang, R. Kondapalli, A. Bachmutsky, **Y. Yuan**. **Offload of Data Lookup Operations**, *US Patent App. 16/207,065, filed Nov. 2018*

Work Experience

- **Microsoft Research** *June 2020 – August 2020*
  - Research Intern at Systems Research Group, Redmond, WA
  - Collaborators: Dan Ports and Jacob Nelson
  - Explored and evaluated new application areas and architectures of modern programmable switch.

• **Intel Labs** *May 2019 – August 2019*
May 2018 – August 2018

– Research Intern at Networking Performance Lab, Hillsboro, OR

– Collaborators: Ren Wang and Yipeng Wang

– Conducted research on next-generation high-performance network platform and I/O system.

Professional Services

• Shadow program committee member at *EuroSys'22*

Research Experience

• **Fine-grained Accelerator Design for Network/Application Dataplane Operations** *2018 – Present*
Tackling the "datacenter tax" problem and the "killer microsecond" problem, We design accelerator architecture, programming models, and integration schemes to accelerate a wide range of fine-grained but costly operations in datacenter's software stacks and applications. The results have been published in ***HPCA'21*** and ***ISCA'19***.

• **I/O Subsystem Design and Optimization for Modern Server CPU** *2018 – 2021*
High-speed I/O devices can exert significant pressure on the CPU's cache/memory system. We study the I/O-host interaction behavior in the real system, and build realistic and accurate I/O subsystem models for gem5 simulator. We also propose multiple solutions in both real systems and simulation models to optimize the data transfer, notification, and interference in the I/O subsystem. **The results have been published in *ISCA'21* and *ISPASS'20*.**

• **In-network Computing for Distributed ML Training Acceleration** *2017 – 2021*
Distributed ML training is notoriously time- and resource-consuming. We propose to leverage the networking devices, including NICs (for in-network gradient compression) and switches (for in-network gradient aggregation), to facilitate the inter-machine communication, which is the most expensive portion in distributed training. We also explore the new potential for P4 programmable switch to process more complicated (floating-point) operations. The results have been published in ***NSDI'22***, ***ISCA'19***, and ***MICRO'18***.

Teaching Experience

• *ECE 411:* Computer Organization and Design (SP 2021)

Skills and Techniques

• *Programming languages:* C/C++, Verilog HDL, VHDL, Python, P4, Shell script, LaTeX, Matlab, etc.

• *Development skills:* Unix/Linux, FPGA, DPDK, RDMA, programmable switch, CUDA, gem5 simulator, sniper simulator, etc.

Selected Courses

• Computer Architecture; High-speed and Programmable Networks; Advanced Memory and Storage System; Distributed System; Advanced Computer Networks; Applied Parallel Programming; Computer Security; System-on-Chip Design; Introduction to VLSI Design; Digital System Design; Embedded System; Artificial Intelligence

Awards and Honors

• OSDI'21 Student Travel Grant *2021*

• NSDI'20 Student Travel Grant *2020*

• OSDI'18 Student Travel Grant *2018*

• Scholarship for Academic Excellence *2016*

• Third Prize in University Robot Contest *2016*

• Scholarship for Academic Excellence *2015*