Yifan Yuan

Email: yifan.yuan@intel.com Website: https://yifanyuan3.github.io

Research Interests

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

Education

- **University of Illinois at Urbana-Champaign** *August 2017 – May 2022*
  - M.S. (2019), Ph.D (2022) in Computer Engineering
  - Advisor: Prof. Nam Sung Kim
  - Doctoral committee: Prof. Nam Sung Kim, Dr. Ren Wang, Prof. Radhika Mittal, Prof. Deming Chen
- **Zhejiang University** *September 2014 – June 2018*
  - B.E. in Electronic Information Engineering

Work Experience

- **Intel Labs** *July 2022 – Present*
  - Research Scientist at Networking Platforms Lab, Hillsboro, OR
- **Microsoft Research** *June 2020 – August 2020*
  - Research Intern at Systems Research Group, Redmond, WA
- **Intel Labs** *May 2019 – August 2019*
  May 2018 – August 2018
  - Research Intern at Networking Platforms Lab, Hillsboro, OR

Publications

- **STYX: Exploiting SmartNIC Capability to Reduce Datacenter Memory Tax**
  H, Ji, Y. Sun, M. Mansi, **Y. Yuan**, J. Huang, R. Kuper, M. Swift, N. S. Kim
  The USENIX Annual Technical Conference (**ATC**), 2023
- **RAMBDA: RDMA-driven Acceleration Framework for Memory-intensive us-scale Datacenter Applications**
  Y. Yuan, J. Huang, Y. Sun, T. Wang, J. Nelson, D. Ports, Y. Wang, R. Wang, C. Tai, N. S. Kim
  The IEEE International Symposium on High-Performance Computer Architecture (**HPCA**), 2023
- **IDIO: Network-Driven, Inbound Network Data Orchestration on Server Processors**
  M. Alian, S. Agarwal, J. Shin, N. Patel, **Y. Yuan**, D. Kim, R. Wang, N. S. Kim
  The ACM/IEEE International Symposium on Microarchitecture (**MICRO**), 2022
- **Unlocking the Power of Inline Floating-Point Operations on Programmable Switches**
  Y. Yuan, O. Alama, J. Fei, J. Nelson, D. R. K. Ports, A. Sapio, M. Canini, N. S. Kim
  The USENIX Symposium on Networked Systems Design and Implementation (**NSDI**), 2022
- **Don't Forget the I/O When Allocating Your LLC**
  Y. Yuan, M. Alian, Y. Wang, R. Wang, I. Kurakin, C. Tai, N. S. Kim
  The ACM/IEEE International Symposium on Computer Architecture (**ISCA**), 2021
  Code to appear in Intel official RDT (pqos) library
- **QEI: Query Acceleration Can be Generic and Efficient in the Cloud**
  Y. Yuan, Y. Wang, R. Wang, R. Chowdhury, C. Tai, N. S. Kim
  The IEEE International Symposium on High-Performance Computer Architecture (**HPCA**), 2021
- **Data Direct I/O Characterization for Future I/O System Exploration**
  M. Alian, **Y. Yuan**, J. Zhang, R. Wang, M. Jung, N. S. Kim
  The IEEE International Symposium on Performance Analysis of Systems and Software (**ISPASS**), 2020
- **HALO: Accelerating Flow Classification for Scalable Packet Processing in NFV**
  Y. Yuan, Y. Wang, R. Wang, J. Huang
  The ACM/IEEE International Symposium on Computer Architecture (**ISCA**), 2019
- **Accelerating Distributed Reinforcement Learning with In-Switch Computing**
  Y. Li, I. Liu, **Y. Yuan**, D. Chen, A. Schwing, J. Huang
  The ACM/IEEE International Symposium on Computer Architecture (**ISCA**), 2019

- **Project Almanac: A Time-Traveling Solid-State Drive**
  X. Wang, **Y. Yuan**, Y. Zhou, C. C. Coats, J. Huang
  The ACM European Conference on Computer Systems (**EuroSys**), 2019
- **A Network-Centric Hardware/Algorithm Co-Design to Accelerate Distributed Training of Deep Neural Networks**
  Y. Li, J. Park, M. Alian, **Y. Yuan**, Q. Zheng, P. Pan, R. Wang, A. Schwing, H. Esmaeilzadeh, N. S. Kim
  The ACM/IEEE International Symposium on Microarchitecture (**MICRO**), 2018

Patents

- **Hardware Assisted Lookup Operations**
  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
  US Patent App. 63/130,663, filed Dec. 2020
- **Data Consistency and Durability over Distributed Persistent Memory Systems**
  R. Wang, **Y. Yuan**, Y. Wang, T.-Y. C. Tai, T. Hurson
  US Patent App. 62/986,094, filed Aug. 2020
- **Workload Scheduler for Memory Allocation**
  Y. Wang, R. Wang, T.-Y. C. Tai, **Y. Yuan**, P. Pathak, S. Vedantham, C. Macnamara
  US Patent App. 16/799,745, filed Feb. 2020
- **Offload of Data Lookup Operations**
  R. Wang, A. J. Herdrich, T.-Y. C. Tai, Y. Wang, R. Kondapalli, A. Bachmutsky, **Y. Yuan**
  US Patent App. 16/207,065, filed Nov. 2018

Professional Service

- **Program Committee:** HPCA (2023, ERC), EuroSys (2022, shadow PC)
- **Reviewer:** IEEE Transactions on Parallel and Distributed Systems (TPDS, 2022), IEEE Computer Architecture Letter (CAL, 2022)

Research Experience

- **Accelerator Design for Network/Application Dataplane Operations** *2018 – Present*
  UIUC and Intel Labs
  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*
  UIUC and Intel Labs
  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 ***MICRO'22***, ***ISCA'21*** and ***ISPASS'20***.
- **In-network Computing for Distributed ML Training Acceleration** *2017 – 2021*
  UIUC and Microsoft Research
  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

- Student Travel Grant: NSDI'22, OSDI'21, HPCA'21, NSDI'20, OSDI'18