

# Yifan Yuan

Email: yifany3@gmail.com

Website: <https://yifanyuan3.github.io>

## Research Interests

---

- Networking hardware and system software for datacenter
- Hardware-software co-design for distributed/disaggregated systems acceleration
- Modern heterogeneous memory system and software

## 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
- **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, Santa Clara, CA
- **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

---

- **Nomad: Non-Exclusive Memory Tiering via Transactional Page Migration**  
L. Xiang, Z. Lin, W. Deng, H. Lu, J. Rao, **Y. Yuan**, R. Wang  
*The USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, 2024
- **Intel Accelerator Ecosystem: An SoC-Oriented Perspective**  
**Y. Yuan**, R. Wang, N. Ranganathan, N. Rao, S. Kumar, P. Lantz, V. Sanjeevan, J. Cabrera, A. Kwatra, R. Sankaran, I. Jeong, N. S. Kim  
*The ACM/IEEE International Symposium on Computer Architecture (ISCA, industry track)*, 2024  
**Paper based on real Intel product and software ecosystem**
- **A Quantitative Analysis and Guidelines of Data Streaming Accelerator in Modern Intel Xeon Scalable Processors**  
R. Kuper, I. Jeong, **Y. Yuan**, R. Wang, N. Ranganathan, N. Rao, J. Hu, S. Kumar, P. Lantz, N. S. Kim  
*The ACM Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, 2024  
**Paper based on real Intel product and software ecosystem**
- **BonsaiKV: Towards Fast, Scalable, and Persistent Key-Value Stores with Tiered, Heterogeneous Memory System**  
M. Cai, J. Shen, **Y. Yuan**, Z. Qu, B. Ye  
*The International Conference on Very Large Databases (VLDB)*, 2024
- **Demystifying CXL Memory with Genuine CXL-Ready Systems and Devices**  
Y. Sun, **Y. Yuan**, Z. Yu, R. Kuper, C. Song, J. Huang, H. Ji, S. Agarwal, J. Lou, I. Jeong, R. Wang, J. H. Ahn, T. Xu, N. S. Kim  
*The ACM/IEEE International Symposium on Microarchitecture (MICRO)*, 2023  
**The first CXL memory paper based on real CXL systems**  
**Featured in Semiconductor Engineering**
- **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 merged into 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

---

- **Method and Apparatus for Batching Pages for a Data Movement Accelerator**  
R. Wang, **Y. Yuan**, R. Kuper  
*US Patent App. 18/477,628*, filed Sep. 2023
- **Efficiently Merging Non-identical Pages in Kernel Same-page Merging (KSM) for Efficient and Improved Memory Deduplication and Security**  
R. Kuper, R. Wang, **Y. Yuan**  
*US Patent App. 18/369,090*, filed Sep. 2023
- **Data Consistency and Durability over Distributed Persistent Memory Systems**  
R. Wang, **Y. Yuan**, Y. Wang, T.-Y. C. Tai, T. Hurson  
*US Patent 11,709,774*, granted Jul. 2023
- **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 11,698,929*, granted Jul. 2023

## Professional Services and Activities

---

- **Live Demo Presenter:** Improve System Performance by Offloading Memory-Intensive Kernel Features to CXL Type-2 Device (OCP Global Summit 2023)
- **Tutorial Organizer and Presenter:** On-chip Accelerators in 4th Gen Intel® Xeon® Scalable Processors: Features, Performance, Use Cases, and Future! (ISCA'2023)
- **Program Committee:** ISCA'2024 (ERC and industry track committee), HPCA'2024, HPCA'2023 (ERC), EuroSys'2022 (shadow PC)
- **Reviewer:** IEEE Transactions on Parallel and Distributed Systems (TPDS, 2022), IEEE Computer Architecture Letter (CAL, 2022-2023), ACM Transactions on Architecture and Code Optimization (TACO, 2024)

## Research Experience

---

- **Datacenter and System Taxes Reduction**

2023 – Present

*Intel Labs*

The concept of the “datacenter and systems taxes” refer to a set of shared low-level software components that consume a significant portion of processor cycles in warehouse-scale computers (WSCs) and distributed systems. We have been conducting various hardware and software innovations to reduce such taxes. The results have been published in *ISCA’24*, *ASPLOS’24*, *OCP 2023 Global Summit*, and *USENIX ATC’23*, and have been transferred to Intel (future) products.

- **Embracing Emerging CXL devices in Modern Datacenter**

2022 – Present

*Intel Labs*

CXL has been attracting much attention as the next generation of device interconnect standard, providing unique features such as memory expansion and cache coherence. **As the pioneer of this direction**, we have been exploring CXL memory devices and CXL accelerators based on real commodity hardware from different aspects, including both hardware functions enhancements and system software optimizations. The results have been published in *OSDI’24*, *VLDB’24*, *MICRO’23* and *OCP 2023 Global Summit*.

- **Accelerator Design for Network/Application Dataplane Operations**

2018 – 2023

*UIUC and Intel Labs*

Tackling the modern data explosion and the “killer microsecond” problem in datacenters, 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’23*, *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 (UIUC, 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.