YANPENG YU

New Haven, Connecticut | (475) 280-1045 | yanpeng.yu@yale.edu | Personal Website: https://yanpeng-yu.com/

EDUCATION

Yale University, New Haven, CT, USA 09/2021– 05/2026 (expected)
Ph.D. in Computer Science
Advisors: Profs. Anurag Khandelwal and Lin Zhong
Peking University, Beijing, China 09/2017 – 06/2021
B.S. in Computer Science
Exchange program: Stanford University, 2019 Summer

AREA OF EXPERTISE

- Computer systems, computer architectures, distributed systems, operating systems, memory disaggregation, cache coherence protocols, synchronization, networks, system and architecture for AI/ML/LLM
- C/C++, CUDA, Python, PyTorch, vLLM, Linux, RDMA

WORK EXPERIENCE

NVIDIA Corporation, Architecture Research Group, *Research Intern*, Santa Clara, CA 05/2025 – 08/2025

- Multi-GPU load-balancing for efficient expert-parallel mixture-of-expert (MoE) model serving.
- Reduced Qwen3 and DeepSeek-V3 decode latency by up to 22% and improved vLLM token throughput by up to 21% compared to DeepSeek's EPLB.
- Paper in submission to ISCA '26.

NVIDIA Corporation, Architecture Research Group, *Research Intern*, Santa Clara, CA 05/2024 – 08/2024

- High-performance and energy-efficient cache coherence protocols for CPU-GPU shared memory.
- Reduced CPU-GPU collaborative benchmarks' running time by 24% and inter-PU communication traffic by 13% compared to existing heterogeneous cache coherence protocols.
- Paper accepted to ISCA '25, won distinguished artifact award; US patent application filed.

SELECTED PUBLICATIONS AND PREPRINTS

- **Yanpeng Yu**∗, Haiyue Ma*, (*equal contribution) Krish Agarwal, Nicolai Oswald, Qijing Huang, Hugo Linsenmaier, Chunhui Mei, Ritchie Zhao, Ritika Borkar, Bita Rouhani, David Nellans, Ronny Krashinsky, Anurag Khandelwal. Efficient MoE Serving in the Memory-Bound Regime: Balance Activated Experts, Not Tokens. *Under Review, 2025.*
- **Yanpeng Yu,** Nicolai Oswald, and Anurag Khandelwal. CORD: Low-Latency, Bandwidth-Efficient and Scalable Release Consistency via Directory Ordering. *In Proceedings of the 52nd Annual International Symposium on Computer Architecture (ISCA), 2025.* 🏆 **Distinguished Artifact Award**
- **Yanpeng Yu,** Seung-seob Lee, Lin Zhong, and Anurag Khandelwal. GCS: Generalized Cache Coherence For Efficient Synchronization . *Under Review, 2023.*
- Lei Zou, Fan Zhang, Yinnian Lin, and **Yanpeng Yu**. An Efficient Data Structure for Dynamic Graph on GPUs. *IEEE Transactions on Knowledge and Data Engineering (TKDE), 2023.*
- Seung-seob Lee, **Yanpeng Yu,** Yupeng Tang, Anurag Khandelwal, Lin Zhong, and Abhishek Bhattacharjee. MIND: In-Network Memory Management for Disaggregated Data Centers. *In Proceedings of the ACM SIGOPS 28th Symposium on Operating Systems Principles (SOSP), 2021.*

AWARDS

🏆 **Distinguished Artifact Award**, 52nd Annual International Symposium on Computer Architecture (ISCA) 2025

LEADERSHIP EXPERIENCE

Athena Student Leadership Council, Athena AI Institute, *Student Leader* 2024 – 2025

TEACHING EXPERIENCE

- Teaching Assistant, CPSC 438/538 Big Data Systems: Trends & Challenges, Yale University 2023 Fall
- Teaching Assistant, CPSC 437 Introduction to Database Systems, Yale University 2022 Fall