BINGYAO LI

210 S. Bouquet Street, Sennott Square 6504, Pittsburgh, PA, 15232
bil35@pitt.edu +1 (412) 616-5592 libingyao.github.io

EDUCATION

University of Pittsburgh Aug. 2020 - Present
Ph.D. in Computer Science
Advisor: Dr. Xulong Tang

Tianjin University Sep. 2017 - Jan. 2020
M.S. in Computer Science and Technology
Advisor: Dr. Ce Yu, Graduated with Honor

Tianjin University Sep. 2013 - July 2017
B.E. in Computer Science and Technology
Graduated with Honor

PUBLICATIONS

[1] **Bingyao Li**, Yueqi Wang, Tianyu Wang, Lieven Eeckhout, Jun Yang, Aamer Jaleel, Xulong Tang, "STAR: Sub-Entry Sharing-Aware TLB for Multi-Instance GPU", *In Proceedings of the 57th IEEE/ACM International Symposium on Microarchitecture.* **(MICRO 2024)**

[2] Yueqi Wang*, **Bingyao Li***, Aamer Jaleel, Jun Yang, Xulong Tang, "GRIT: Enhancing Multi-GPU Performance with Fine-Grained Dynamic Page Placement", *The 30th IEEE International Symposium on High-Performance Computer Architecture.* **(HPCA 2024)**, * The authors contribute equally.

[3] **Bingyao Li**, Yanan Guo, Yueqi Wang, Aamer Jaleel, Jun Yang, Xulong Tang, "IDYLL: Enhancing Page Translation in Multi-GPUs via Light Weight PTE Invalidations", *In Proceedings of the 56th IEEE/ACM International Symposium on Microarchitecture.* **(MICRO 2023)**

[4] **Bingyao Li**, Yueqi Wang, Xulong Tang, "Orchestrated Scheduling and Partitioning for Improved Address Translation in GPUs", *The 60th Design Automation Conference.* **(DAC 2023)**

[5] **Bingyao Li**, Jieming Yin, Anup Holey, Youtao Zhang, Jun Yang, Xulong Tang, "Trans-FW: Short Circuiting Page Table Walk in Multi-GPU Systems via Remote Forwarding", *The 29th IEEE International Symposium on High-Performance Computer Architecture.* **(HPCA 2023)**

[6] **Bingyao Li***, Qi Xue*, Geng Yuan*, Sheng Li, Xiaolong Ma, Yanzhi Wang and Xulong Tang, "Optimizing Data Layout for Training Deep Neural Networks", *The ACM Web Conference Workshop.* **(WWW 2022)**, * The authors contribute equally.

[7] **Bingyao Li**, Jieming Yin, Youtao Zhang, Xulong Tang, "Improving Address Translation in Multi-GPUs via Sharing and Spilling aware TLB Design", *In Proceedings of the 54th IEEE/ACM International Symposium on Microarchitecture.* **(MICRO 2021)**

[8] **Bingyao Li**, Ce Yu, Chen Li, Xiaoteng Hu, Jian Xiao, Shanjiang Tang, Chenzhou Cui, and Dongwei Fan, "mcatCS: A Highly Efficient Cross-Matching Scheme for Multi-Band Astronomical Catalogs", *Publication of the Astronomical Society of the Pacific, 2019, 131(999).*

[9] Ce Yu, **Bingyao Li**, Jian Xiao, Chao Sun, Shanjiang Tang, Chongke Bi, Chenzhou Cui, and Dongwei Fan, "Astronomical Data Fusion: Recent Progress and Future Prospects - A Survey", *Springer Experimental Astronomy, 2019(6).*

[10] **Bingyao Li**, Ce Yu, Xiaoteng Hu, Jian Xiao, Shanjiang Tang, Lianmeng Li, Bin Ma, "An Efficient Retrieval Method for Astronomical Catalog Time Series Data", *The 18th International Conference on Algorithms and Architectures for Parallel Processing* **(ICA3PP 2018)**

[11] Xiaoteng Hu, Ce Yu, **Bingyao Li**, Shanjiang Tang, Jian Xiao, Yanyan Huang, "GAIDR: An Efficient Time Series Subsets Retrieval Method for Geo-Distributed Astronomical Data", *The 20th IEEE International Conference on High Performance Computing and Communications* **(HPCC 2018)**

RESEARCH EXPERIENCE

NVIDIA Research May 2024 - Aug. 2024
Research Intern, Architecture Research Group
Mentor: Dr. Aamer Jaleel
- Develop KV-cache hierarchy for multi-turn interaction LLM inference system
- Explore optimization space of transferring KV-cache between GPU memory and host DRAM systems

University of Pittsburgh Aug. 2020 - Present
Research Assistant
Advisor: Dr. Xulong Tang
- Design architectures and system features for multi-GPU systems, with a focus on virtual memory
- Develop flexible and reconfigurable GPUs for multi-tenant execution
- Develop efficient data layout management for deep learning application

Tianjin University Sep. 2017 - Jan. 2020
Research Assistant
Advisor: Dr. Ce Yu
- Develop time series subsets retrieval system for large-scale astronomical image data
- Optimize cloud-based storage for long-term astronomical archive data
- Develop distributed cross-matching scheme for billion-row astronomical data
- Design automatic method for cross-matching celestial objects accurately

ICT of Chinese Academy of Science June 2019 - Aug. 2019
Visiting Scholar
Advisor: Dr. Yungang Bao
- Port latency-sensitive benchmark to RISC-V architecture
- Evaluate the performance of Tailbench-Riscv on LvNA (Labeled RISC-V)

SELECTED HONORS & AWARDS

Student Travel Grant, HPCA	2023, 2024
Selected PhD Forum Attendee at MICRO	2023
CS50 Outstanding Research Fellowship, University of Pittsburgh	2022, 2023
Student Travel Grant, MICRO	2022, 2023
Student Travel Grant, ISCA	2022
SCI Fellowship, University of Pittsburgh	2020
National Scholarship, Ministry of Education of China	2019
Graduate Scholarship - First Prize, Tianjin University	2017, 2019

RESEARCH TALKS

- **Towards Efficient and Salable Computing for Multi-GPUs** 2024
  at NVIDIA, USA
- **GRIT: Enhancing Multi-GPU Performance with Fine-Grained Dynamic Page Placement** 2024
  at HPCA 2024, Edinburgh, UK

- **IDYLL: Enhancing Page Translation in Multi-GPUs via Light Weight PTE Invalidations** 2023
  at MICRO 2023, Toronto, ON
- **Orchestrated Scheduling and Partitioning for Improved Address Translation in GPUs** 2023
  at DAC 2023, San Francisco, CA
- **Towards Efficient and Salable Computing for Multi-GPUs** 2023
  at Tianjin University, China
- **Trans-FW: Short Circuiting Page Table Walk in Multi-GPU Systems via Remote Forwarding** 2023
  at HPCA 2023, Montreal, QC
- **Optimizing Data Layout for Training Deep Neural Networks** 2022
  at WWW 2022, Virtual
- **Improving Address Translation in Multi-GPUs via Sharing and Spilling aware TLB Design** 2021
  at MICRO 2021, Virtual

TEACHING

- Teaching Assistant of CS 1550: Introduction to Operating Systems, Pitt, Fall 2021

PROFESSIONAL SERVICE

Artifact Evaluation Committee of MICRO'22, ASPLOS'23, MICRO'24