BINGYAO LI

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

EDUCATION

Aug. 2020 – Now	**University of Pittsburgh** Ph.D. in Computer Science Advisor: Dr. Xulong Tang
Sep. 2017 – Jan. 2020	**Tianjin University** M.S. in Computer Science and Technology Advisor: Dr. Ce Yu, Graduated with Honor
Sep. 2013 – July 2017	**Tianjin University** B.E. in Computer Science and Technology Graduated with Honor

RESEARCH EXPERIENCE

May 2024 - Aug. 2024	**NVIDIA Research**, *Architecture Research Group* Research Intern working on KV-cache management for LLM inference Mentor: Dr. Aamer Jaleel
June 2019 - Aug. 2019	**ICT of Chinese Academy of Science** Visiting Scholar working on benchmarking RISC-V architecture Mentor: Dr. Yungang Bao

CONFERENCE PUBLICATIONS

* The authors contribute equally.

HPCA 2025 Yueqi Wang, **Bingyao Li**, Mohamed Tarek Ibn Ziad, Lieven Eeckhout, Jun Yang, Aamer Jaleel, Xulong Tang, "OASIS: Object-Aware Page Management for Multi-GPU Systems", *The 31th IEEE International Symposium on High-Performance Computer Architecture*, 2025.

MICRO 2024 **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*, 2024.

HPCA 2024 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*, 2024.

MICRO 2023 **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*, 2023.

HPCA 2023 **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*, 2023.

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

WWW 2022	**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*, 2022.
MICRO 2021	**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*, 2021.
ICA3PP 2018	**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*, 2018
HPCC 2018	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*, 2018

JOURNAL

PASP 2019	**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).
EA 2019	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).

PATENTS

2024	**Bingyao Li**, Aamer Jaleel, Po-An Tsai, Anish Saxena, "Dynamic Key-Value Cache Scheduling for LLM Serving", NVIDIA (pending)
2024	Anish Saxena, Po-An Tsai, **Bingyao Li**, Aamer Jaleel, "Restricted Expert Mixture Framework for MoE LLM Serving", NVIDIA (pending)

HONORS & AWARDS

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

RESEARCH TALKS

2024	**STAR: Sub-Entry Sharing-Aware TLB for Multi-Instance GPU** at MICRO 2024, Austin, TX
2024, 2023	**Towards Efficient and Salable Computing for Multi-GPUs** at NVIDIA, USA at Tianjin University, China

2024	**GRIT: Enhancing Multi-GPU Performance with Fine-Grained Dynamic Page Placement** at HPCA 2024, Edinburgh, UK
2023	**IDYLL: Enhancing Page Translation in Multi-GPUs via Light Weight PTE Invalidations** at MICRO 2023, Toronto, ON
2023	**Orchestrated Scheduling and Partitioning for Improved Address Translation in GPUs** at DAC 2023, San Francisco, CA
2023	**Trans-FW: Short Circuiting Page Table Walk in Multi-GPU Systems via Remote Forwarding** at HPCA 2023, Montreal, QC
2022	**Optimizing Data Layout for Training Deep Neural Networks** at WWW 2022, Virtual
2021	**Improving Address Translation in Multi-GPUs via Sharing and Spilling aware TLB Design** at MICRO 2021, Virtual

ACADEMIC SERVICES

Program Committee	The International Conference on Learning Representations (ICLR 2025)
Journal Reviewer	IEEE Transactions on Computers (TC 2024) ACM Transactions on Architecture and Code Optimization (TACO 2024) Journal of Combinatorial Optimization (JOCO 2024)
Artifact Evaluation Committee	IEEE/ACM International Symposium on Microarchitecture (MICRO 2024, 2022) ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS 2023)

TEACHING EXPERIENCE

Fall 2023, 2022	Guest Lecturer for Computer Architecture, CS 2410, UPitt
Fall 2021	Teaching Assistant of Introduction to Operating Systems, CS 1550, UPitt

MENTORING EXPERIENCE

Fall 2022 - now	Yueqi Wang (University of Pittsburgh, Ph.D. student) Page placement in multi-GPU, **published paper at HPCA'24, HPCA'25**
Fall 2022 - now	Tianyu Wang (University of Pittsburgh, Ph.D. student) ML on multi-instance GPU, submitted a paper to ISCA'25
Fall 2024 - now	Ziwei Quan (University of Pittsburgh, Master student) Ray tracing accelerator

REFERENCE

Dr. Xulong Tang
Assistant Professor
University of Pittsburgh
tax6@pitt.edu

Dr. Aamer Jaleel
Principal Research Scientist
NVIDIA
ajaleel@nvidia.com

Dr. Jun Yang
Professor
University of Pittsburgh
juy9@pitt.edu

Dr. Lieven Eeckhout
Professor
Ghent University
Lieven.Eeckhout@ugent.be

Dr. Bruce R. Childers
Professor & Dean of the School of Computing and Information
University of Pittsburgh
childers@pitt.edu