# Zihao Ye

Ph.D. student

*Bill & Melinda Gates Center #330*
*185 NE Stevens Way, Seattle, WA, 98105*
*(206)-228-6099*
*zhye@cs.washington.edu*
*homes.cs.washington.edu/ zhye*
*yzh119*


## Research Interest

I have broad interests in Computer Systems, Compiler, Programming Languages, and Computer Architecture. My current research centers around sparse computation:

- Programming Abstractions for Sparse Workloads on Heterogeneous Hardware.
- Domain Specific Accelerator and Memory Architecture for Sparsity.
- Scalable and Affordable Sparse Computation in Data Centers.

## Education

2021 Spring – Present **Ph.D. student in Computer Science**, *University of Washington*.
SAMPL(System, Architecture, Machine learning, and Programming language) Group
Adviser: Luis Ceze

2014–2018 **B.Eng in Computer Science**, *Shanghai Jiao Tong University*.
Member of ACM Honors Class 2014.

## Employment

2021 **MLSys Engineer Intern**, *OctoML*, Seattle, WA.

2018–2021 **Software Development Engineer II**, *Amazon Web Services*, Shanghai, China.

2017–2018 **Research Intern**, *Microsoft Research Asia*, Beijing, China.

## Publications

ASPLOS 2023 Zihao Ye, Ruihang Lai, Junru Shao, Tianqi Chen, and Luis Ceze. SparseTIR: Composable Abstractions for Sparse Compilation in Deep Learning. *The 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*, 2023.

ASPLOS 2023 Siyuan Feng, Bohan Hou, Hongyi Jin, Wuwei Lin, Junru Shao, Ruihang Lai, Zihao Ye, Lianmin Zheng, Cody Hao Yu, Yong Yu, and Tianqi Chen. TensorIR: An Abstraction for Automatic Tensorized Program Optimization. *The 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*, 2023.

MLSys 2022 Zhiqiang Xie, Minjie Wang, Zihao Ye, Zheng Zhang, and Rui Fan. Graphiler: Optimizing Graph Neural Networks with Message Passing Data Flow Graph. In *Proceedings of Machine Learning and Systems*, 2022.

SC 2020 Yuwei Hu, Zihao Ye, Minjie Wang, Jiali Yu, Da Zheng, Mu Li, Zheng Zhang, Zhiru Zhang, and Yida Wang. FeatGraph: A Flexible and Efficient Backend for Graph Neural Network Systems. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, 2020.

Preprint Zihao Ye, Qipeng Guo, Quan Gan, Xipeng Qiu, and Zheng Zhang. BP-Transformer: Modelling Long-Range Context via Binary Partitioning. *arXiv preprint arXiv:1911.04070*, 2019.

Preprint Minjie Wang, Da Zheng, Zihao Ye, Quan Gan, Mufei Li, Xiang Song, Jinjing Zhou, Chao Ma, Lingfan Yu, Yu Gai, Tianjun Xiao, Tong He, George Karypis, Jinyang Li, and Zheng Zhang. Deep Graph Library: A Graph-Centric, Highly-Performant Package for Graph Neural Networks. *arXiv preprint arXiv:1909.01315*, 2019.

## Invited Talks

### SparseTIR

Nov 2022 Amazon AI

Oct 2022 CRISP Liaison Meeting

Aug 2022 NICS-EFC Lab, Tsinghua University

Aug 2022 Zhang Research Group, Cornell University

July 2022 MLIR Reading Group, Google

Dec 2021 TVMCon (Apache TVM and Open Source ML Acceleration Conference)

## Awards and Grants

2015 Gold Medal in 2015 ACM-ICPC China Shanghai Metropolitan Programming Contest

## Academic Services

External Review Committee MLSys 2023

Artifact Evaluation Committee MLSys 2022, CGO 2023

Reviewer IEEE Computer Architecture Letters

Organizer UW SAMPL Seminar

## Open Source Activity

Commiter DGL(Deep Graph Library)

Reviewer Apache TVM