Zhisheng YE Ph.D. Student

School of Computer Science, Peking University
1336, 1st Science Building, No.5 Yiheyuan Lu
Haidian District, Beijing, 100871, P. R. China
yezhisheng@pku.edu.cn
https://yezhisheng.me · yzs981130 · yzs981130

Research Interests

I am interested in Distributed Systems, Systems for Machine Learning and Resource Management. My current focuses include:

- Scheduling systems for Deep Learning training workloads.
- Elastic & flexible resource management on heterogeneous clusters.
- Co-optimizing deep learning framework with scheduling.

Education

Peking University Beijing, China
Pursuing Ph.D. in Computer Science *Sep 2019 – Present*
Related Courses: Distributed Systems / Introduction to Parallel Computing / Dataflow Algorithms
Advisor: Prof. Yingwei Luo & Prof. Xiaolin Wang

Peking University Beijing, China
Bachelor of Science in Computer Science & Technolegy *Sep 2015 – Jun 2019*
Related Courses: Computer Network (Honor track) / Computer Network Practice (Honor track) / Operating System (Honor track) / Introduction to Computer System

Professional Experience

Shanghai AI Laboratory Beijing, China
Research Intern *July 2022 – Jan 2024*

- Focusing on large scale model training infrastructure optimization. Deeply involved in the development of InternLM, a large language model with over 100 billion parameters.
- Dense LLM Optimization
  - Storage and input data layout optimization for LLM training; 5× speedup over original of data sample loading and 10% improvement of end-to-end training.
  - Automatic profiling and communication hotspot discovery for LLM training.
  - Co-optimize LLM training jobs with scheduling system. Developed an efficient scheduler for colocating LLM training jobs and Hyperparameter Optimization (HPO) jobs.
- Sparse Model (MoE-like) Training Optimization
  - All-to-all communication profiling and optimization for MoE-like models under inter-node communication bottleneck.
  - Implemented dynamic expert loading and unloading for mechanisms for MoE-like models. Overlapped expert loading and unloading with communication and computation to improve throughput.

Sensetime Research Beijing, China
Research Intern *Sept 2019 – June 2022*

- Advisor: Peng Sun from Sensetime Research and Tianwei Zhang from S-Lab, Nanyang Technological University.
- Creativity projects
  - Developed and drafted Astraea, a fair scheduler for deep learning training jobs, as first author.
  - Authored 5 patents related to scheduling as first author.
- Communication optimization of GPUs across pods
  - Modified the logic of kubelet, communication framework and application framework to use GPU, to achieve cross-pod P2P communication
  - Achieved near bare-metal performance; 35% improvement over original.
- Designed and implemented a production scheduler on Kubernetes, introducing GPU topology-aware scheduling.
- Cluster trace analysis and characteristics of deep learning workloads, aiming at pending situation and daily / monthly trends of job behaviors.

Peng Cheng Laboratory Shenzhen, China
Research Intern *July 2018 – Sept 2021*

- Contributed to development of OpenI-Octopus, an open-sourced scheduler for deep learning training workloads based on Kubernetes.
- Safe GPU sharing mechanisms for multiple workloads
  - Intercepted job usage of GPU by deep learning framework analysis and CUDA driver / library wrapping to securely limit job usage of GPU memory.
  - Open-sourced the mechanisms of safe GPU sharing under memory limitation.
  - Supported high-efficient pausing and resuming of deep learning workloads, which is imperceptible to users.
  - Workloads migration and further combination with scheduling algorithms.
- GPU sharing techniques on Kubernetes
  - Implemented resource allocation and usage techniques of vGPU on Kubernetes as a device plugin. The division of GPUs can be pre-defined or set dynamically.
  - Implemented a monitoring system of vGPU runtime statistics based on Prometheus.

Student Super Computing Team of Peking University Beijing, China
Team member *Sept 2018 – June 2019*

- Participated in analyzing, compiling, profiling, optimizing, and improving parallelizability of general HPC tasks.
- Optimized the compilation and operation of a specific supercomputing task: the gene assembly problem wtdbg2 in ASC19 Student Supercomputer Challenge; Participated and led the group competition challenge in the final round.

Skills

Programming C/C++, Golang, Python, Bash, Javascript
Software & Tools PyTorch, Kubernetes, CUDA
Languages English, Mandarin

Publications

1. Q. Hu, **Z. Ye**, Z. Wang, *et al.*, "Characterization of large language model development in the datacenter," in *21st USENIX Symposium on Networked Systems Design and Implementation (**NSDI '24**)*, 2024, *First author with equal contributions.*
2. **Z. Ye**, W. Gao, Q. Hu, P. Sun, X. Wang, Y. Luo, T. Zhang, and Y. Wen, "Deep learning workload scheduling in gpu datacenters: Taxonomy, challenges and vision," *ACM Computing Surveys*, 2024, *First author with equal contributions.*
3. W. Gao, **Z. Ye**, P. Sun, T. Zhang, and Y. Wen, "Unisched: A unified scheduler for deep learning training jobs with different user demands," *IEEE Transactions on Computers*, 2024.
4. Q. Hu, **Z. Ye**, M. Zhang, Q. Chen, P. Sun, Y. Wen, and T. Zhang, "Hydro: Surrogate-Based hyperparameter tuning service in datacenters," in *17th USENIX Symposium on Operating Systems Design and Implementation (**OSDI '23**)*, 2023.
5. Z. Yang, **Z. Ye**, T. Fu, J. Luo, X. Wei, Y. Luo, X. Wang, Z. Wang, and T. Zhang, "Tear up the bubble boom: Lessons learned from a deep learning research and development cluster," in *2022 IEEE 40th International Conference on Computer Design (**ICCD '22**)*, 2022.
6. **Z. Ye**, P. Sun, W. Gao, T. Zhang, X. Wang, S. Yan, and Y. Luo, "Astraea: A fair deep learning scheduler for multi-tenant gpu clusters," *IEEE Transactions on Parallel and Distributed Systems (**TPDS**)*, 2021.
7. W. Gao, **Z. Ye**, P. Sun, Y. Wen, and T. Zhang, "Chronus: A novel deadline-aware scheduler for deep learning training jobs," in *Proceedings of the ACM Symposium on Cloud Computing (**SoCC '21**)*, Association for Computing Machinery, 2021.

Teaching

- Teaching Assistant: Introduction to Computing (Peking University) *2019 Fall, 2020 Fall, 2021 Fall*
- Teaching Assistant: Introduction to Virtualization and Storage Systems (Peking University) *2019 Fall*

Awards

- Award for Scientific Research, Peking University *2022, 2023*
- First Price (Team), ASC Student Supercomputer Challenge *2019*
- Outstanding Winner, Microsoft Student Club Practice Space *2018*