

YUANXIN WEI

weiyx25@mail2.sysu.edu.cn

EDUCATION

Sun Yat-sen University, Computer Science and Technology, *PH.D* 2023/09 - present

- Lab: National Supercomputing Center in Guangzhou
- Advisor: Prof. Nong Xiao
- Research Interests: Machine learning system and high performance computing

Sun Yat-sen University, Computer Science and Technology, *Bachelor* 2019/09-2023/06

- GPA: 3.94/5.0
- 2023 Outstanding Undergraduate Graduates of Sun Yat-sen University (Top-5%)
- Courses: Operating system, database, computer network, parallel and distributed computing

RESEARCH EXPERIENCE

Fine-tuning MoE Models with Affinity-aware Pipeline Parallelism 2023/09-2024/03

- Design the APTMoE system for fine-tuning MoE models on bandwidth-constrained scenarios.
- Propose a hierarchical loading strategy for computing affinity awareness by strategically offloading a portion of experts to CPU for computation.
- Propose a demand-priority scheduling strategy that dynamically coordinates the loading behaviors, for alleviating mutual interference between different loading phases and maximizing the bandwidth utilization.
- Accepted by SC 2024 conference (22.7%).

Communication-Efficient Distributed Inference for Transformer Models 2023/04-2023/09

- Design a communication-efficient distributed inference system, DeTransformer.
- Adopting the concept of co-design, propose block parallelism through model structure decoupling, incorporated with a model adaptive execution method that dynamically balances the computing power, communication power, and memory capacity of devices.
- Conduct accuracy experiments through pre-training Bert and GPT2 models, and validate their accuracy on downstream tasks; Conduct performance experiments, and achieve 2.81x inference performance improvement on 4 Raspberry Pi devices in an edge bandwidth environment.
- Accepted by DATE 2024 conference (25%).

WORKING EXPERIENCE

Alibaba • PAI • Beijing • Research Intern 2024/07-present

- Conduct research and optimize the training performance of MoE models.

ByteDance • Lark • Shenzhen • Technical Support 2022/05-2022/10

- Collaborate to solve technical problems, and settle issues.

SKILL

- **Coding:** C, C++, Python
- **Tools:** OpenMP, MPI, CUDA, PyTorch, Matlab, L^AT_EX
- **English:** CET-4 and CET-6

AWARDS

- **National scholarship** , Top-1% 2021/12
- **Principal scholarship of Sun Yat-sen University** , Top-5% 2023/09
- **First class scholarship of Sun Yat-sen University** , Top-5% 2021/09
- **Second class scholarship of Sun Yat-sen University ×2** , Top-10% 2020-2022
- **Excellent undergraduate thesis of Sun Yat-sen University** , Top-5% 2023/06

PUBLICATION

- **Yuanxin Wei**, Shengyuan Ye, Jiazhi Jiang, Xu Chen, Dan Huang*, Jiansu Du*, Yutong Lu, *Communication-Efficient Model Parallelism for Distributed In-Situ Transformer Inference*, in Design, Automation & Test in Europe (DATE), 2024, CCF-B.
- **Yuanxin Wei**, Jiansu Du*, Jiazhi Jiang, Xiao Shi, Xianwei Zhang, Dan Huang, Nong Xiao, Yutong Lu*, *APT-MoE: Affinity-aware Pipeline Tuning for MoE Models on Bandwidth-constrained GPU Nodes*, in International Conference for High Performance Computing, Networking, Storage, and Analysis (SC), 2024, CCF-A.