Fahao Chen

✉ chenfh612@gmail.com · ☎ (+86) 134-8707-5757 · in ResearchGate

Profile

My name is Fahao Chen. I received my Ph.D. degree from the Graduate School of Computer Science and Engineering, The University of Aizu in September 2024, supervised by Prof. Peng Li. My research interests include edge computing and distributed machine learning systems.

Education

The University of Aizu, Aizuwakamatsu, Japan 2021 – 2024
PhD student in Computer Science and Information System (CIS)

The University of Aizu, Aizuwakamatsu, Japan 2019 – 2021
Master student in Computer Science and Information System (CIS)

China University of Geosciences (Wuhan), Wuhan, China 2016 – 2020
B.S. in Network Engineering (NE)

Publication

Journal Paper

1. **[IEEE TCC]** Non-Clairvoyant Scheduling of Distributed Machine Learning with Inter-job and Intra-job Parallelism on Heterogeneous GPUs, Fahao Chen, Peng Li, Celimuge Wu, Song Guo, 2024 *(CCF-C)*
2. **[IEEE IoTJ]** Giant Could Be Tiny: Efficient Inference of Giant Models on Resource-Constrained UAVs, Fahao Chen, Peng Li, Shengli Pan, Lei Zhong and Jing Deng, 2024 *(CCF-C)*
3. **[IEEE TPDS]** FedGraph: Federated Graph Learning With Intelligent Sampling, Fahao Chen, Peng Li, Toshiaki Miyazaki and Celimuge Wu, 2022 *(CCF-A)*
4. **[IEEE TCC]** Edge-Assisted Short Video Sharing With Guaranteed Quality-of-Experience, Fahao Chen, Peng Li, Deze Zeng and Song Guo, 2021 *(CCF-C)*

Conference Paper

1. **[IEEE INFOCOM]** SPIN: Accelerating Large Language Model Inference with Heterogeneous Speculative Models, Fahao Chen, Peng Li, Tom H. Luan, Zhou Su, and Jing Deng, 2025 *(CCF-A)*
2. **[IEEE INFOCOM]** Mell: Memory-Efficient Large Language Model Serving via Multi-GPU KV Cache Management, Qianli Liu, Zicong Hong, Peng Li, Fahao Chen, and Song Guo, 2025 *(CCF-A)*
3. **[ACM SIGMOD]** DGC: Training Dynamic Graphs with Spatio-Temporal Non-Uniformity using Graph Partitioning by Chunks, Fahao Chen, Peng Li, and Celimuge Wu, 2024 *(CCF-A)*
4. **[IEEE VTC]** Low-Latency Perception Sharing Services for Connected Autonomous Vehicles, Fahao Chen, Peng Li, Lei Zhong, Dongxiao Yu and Xiuzhen Cheng, 2023 *(CCF-C)*
5. **[ACM ICPP]** TCB: Accelerating Transformer Inference Services with Request Concatenation, Boqian Fu, Fahao Chen, Peng Li, and Deze Zeng, 2022 *(CCF-B)*
6. **[ACM HPDC]** Hare: Exploiting Inter-job and Intra-job Parallelism of Distributed Machine Learning on Heterogeneous GPUs, Fahao Chen, Peng Li, Celimuge Wu, and Song Guo, 2022 *(CCF-B)*

Research Work

- **Efficient Distributed Machine Leaning System:** Distributed machine learning (DML) has shown great promise in accelerating model training on multiple GPUs. To increase GPU utilization, a common practice is to let multiple learning jobs share GPU clusters, where the most fundamental and critical challenge is how to efficiently schedule these jobs on GPUs. However, existing works about distributed machine learning job scheduling are constrained to settings with homogeneous GPUs. GPU heterogeneity is common in practice, but its influence on multiple DML job scheduling has been seldom studied. Moreover, DML jobs have internal structures that contain great parallelism potentials, which have not yet been fully exploited in the heterogeneous computing environment. We propose a novel job scheduler that exploits both inter-job and intra-job parallelism in a heterogeneous GPU cluster, which aims to minimize the average training job completion time.

- **On-Edge Machine Learning Serving System:** Machine Learning (ML) models have demonstrated unprecedented capabilities in handling complex tasks on the edge environment. However, there is an open challenge about the mismatching between the massive computation and memory requirements of ML models and the limited resources on edge devices. Existing works either pose privacy concerns with offloading methods or compromise model accuracy with various model compression techniques. We explore the Mixture-of-Expert (MoE) model architecture that decouples ML models into multiple tiny experts, so that edge devices can dynamically load a few experts that best match their current input.

♡ Honors and Awards

Best Paper Award from IEEE CyberSciTech 2023	Nov. 2023
Best Paper Award from IEEE BDCloud 2023	Dec. 2023
Research Fellowships for Young Scientists from Japan Society for the Promotion of Science	Mar. 2023
Chinese Government Award for Outstanding Self-financed Students Abroad	Aug. 2024