Jin Fang

✉ fanjin98@outlook.com · ☎ (+86) 181-5566-1676 · 🔗 www.fangjin.site

Education

University of Science and Technology of China (USTC) Anhui, China
Ph.D. in Computer Science 2020.9-2026.6 (expected)

- Research focus on MlSys, Collective Communication, and In-network Computing
- Advisors: Prof. Hongli Xu and Prof. Gongming Zhao

Hunan University (HNU) Hunan, China
B.S. in Computer Science 2016.9-2020.6

- Excellent Graduation Thesis of Hunan University

Publications

1. S. Zheng, **J. Fang**, X. Zheng, Q. Hou, W. Bao, N. Zheng, Z. Jiang, D. Wang, J. Ye, H. Lin, L. Chang, X. Liu, *TileLink: Generating Efficient Compute-Communication Overlapping Kernels using Tile-Centric Primitives*, (**MLSys'25**)
2. **J. Fang**, G. Zhao, H. Xu, L. Luo, Z. Yao, A. Xie, *Non-Idle Machine-Aware Worker Placement for Efficient Distributed Training in GPU Clusters*, IEEE International Conference on Network Protocols (**ICNP'24**)
3. **J. Fang**, G. Zhao, H. Xu, Z. Yu, B. Shen, L. Xie, *Accelerating Distributed Training with Collaborative In-network Aggregation*, IEEE/ACM Transactions on Networking (**ToN'24**)
4. **J. Fang**, G. Zhao, H. Xu, Z. Yu, B. Shen, L. Xie, *GOAT: Gradient Scheduling with Collaborative In-Network Aggregation for Distributed Training*, IEEE/ACM International Symposium on Quality of Service (**IWQoS'23**)
5. **J. Fang**, G. Zhao, H. Xu, C. Wu, Z. Yu, *GRID: Gradient Routing with In-network Aggregation for Distributed Training*, IEEE/ACM Transactions on Networking (**ToN'23**)
6. **J. Fang**, G. Zhao, H. Xu, H. Tu, H. Wang, *Reveal: Robustness-Aware VNF Placement and Request Scheduling in Edge Clouds*, Computer Networks (**ComNet'23**)
7. J. Liu, Y. Zhai, G. Zhao, H. Xu, **J. Fang**, Z. Zeng, Y. Zhu, InArt: In-Network Aggregation with Route Selection for Accelerating Distributed Training, International World Wide Web Conference (**WWW'24**)

Experience

Communication Collective Library for Sequence Parallelism LLM Job Bytedance
Seed-Foundation-mlsys, Beijing, China
Main Developer 2024.11-present

- Implement AllGather and AlltoAll operations based on RDMA verbs programming
- Design collective communication algorithms for cross-pcie and cross-node scenarios
- Optimize bandwidth utilization under resource-constraint machine architectures
- Reduce No. of NICs per machine from 8 to 2 (by 75%), while achieve the same bandwidth utilization
- Evaluate performance of different machine architecture and analyze communication bottleneck for different SP setup (SP-Ulysses and SP-Ring)

Communication-Computation Fused GPU Kernel Generation Bytedance
Seed-Foundation-mlsys, Beijing, China
Research Intern 2024.6-2024.12

- Implement collective communication operations (e.g., AllGather) based on Triton and NVSHMEM
- Design and implement communication-computation fused operations (e.g., AllGather+GEMM), exploring overlaps between GEMM and collective communications
- Achieve near-optimized bandwidth utilization on A100*8 NVlink machines
- Implement TP-fused and SP-fused kernels for both dense and MoE LLMs (Integrated into 6 popular LLMs)

- Implement and optimize cross-node communication computation fusion kernels
- Achieve end-to-end speed up by 3× and 1.5× compared with Pytorch and vLLM

Optimizing Worker Placement for Distributed Training in OCS Network Huawei 2012 Lab, Hefei, China

Research Intern 2023.12-2024.5

- Investigate existing large model task deployment and resource scheduling works
- Investigate existing gradient compression optimization for sparse model training
- Model physical and logical communication patterns of different collective communication algorithms, analyze the impact of communication topology on task training time
- Design a task placement algorithm to optimize the cross-rack traffic in the optical circuit switch network

Simulating network faults with programmable dataplane Suzhou, China

Main Developer 2022.12-2023.9

- Build a user-friendly, multi-backend fault injection system in programmable dataplane
- Design a parser generation algorithm to handle flow dependency and load the table entries
- Formulate the fault injection point selection problem
- Implement several network faults with P4 in TNA and PSA architectures

Accelerating distributed training with programmable switches Zhijiang Lab, Hangzhou, China

Research Intern 2022.6-2022.9

- Improve the in-network aggregation throughput by mitigating the influence of asychronous arrived packets
- Design a knapsack-based randomized rounding algorithm for gradient scheduling
- Implement a distributed training prototype with Pytorch
- Implement the in-network aggregation logic in Tofino with P4
- Reduce the communication overhead of distibuted training tasks by 81.2%

Developing and testing Alcor, a cloud native SDN platform Futurewei, *Remotely*

Developer 2021.6-2021.9

- Write an automatic building script for large scale deployment with bash
- Write an end-to-end test of the virtualization control plane (ACA) with C++
- Develop grpc thread for pulsar subscribe information (PR #274) with C++

Implement a LSTM model based on high-level synthesis Hunan, China

Main Developer 2019.6-2020.1

- Train a LSTM model based on Keras to predict the steam pressure in nuclear power plant reactor
- Implement the trained LSTM model with C++ and deploy it into a Pynq-Z2 board
- Reduce the inference time by 90x compared with software implementation
- *Win the award of Excellent Graduation Thesis of Hunan University*

Awards

• Guorui scholarship	2023
• Excellent price (25%) in Intel P4 China Hackthon	2022
• Doctoral first-class academic scholarship	2022, 2023
• Master's first-class study scholarship	2020, 2021

Skills

- Programming Language: C/C++, Python, P4, C#, Swift
- Developing Framework: Pytorch, p4c, eBPF, Mininet