Jiarui Fang (方佳瑞)

+(86) 13717819702 fangjiarui123@gmail.com
https://github.com/feifeibear (200+ followers)
Update Date: July 27, 2022

Extensive research and engineering experience in developing high-performance computing systems. Core contributor to a variety of open-source software that have earned over **6k stars** on GitHub. First author of publications on top-tier system conferences.

Work Experience

- **HPC-AI TECH, a startup** Beijing, China
  Co-founder Engineer, Tech Manager *February 2022 - Now*
- **WeChat, Tencent** Beijing, China
  Senior Software Engineer (T11) *at WeChat AI* *July 2019 - February 2022*
  Mentor : Dr. Jie Zhou and Dr. Cheng Niu
- **National Supercomputing Center in Wuxi** Wuxi, Jiangsu, China
  Ph.D Research Intern *at R&D Center* *August 2017 - May 2019*
  Mentor : Prof. Haohuan Fu

Education

- **Tsinghua University** Beijing, China
  Ph.D. *in Department of Computer Science & Technology* *July 2019*
  Advisor : Prof. Guangwen Yang, Co-advisor: Prof. Haohuan Fu
  Dissertation: Parallel Deep Learning Training System on on Sunway TaihuLight [pdf]
- **University of California, Davis** Davis, CA, USA
  Visiting Scholar *in Department of Computer Science Engineering* *August 2017 - August 2018*
  Advisor : Assistant Prof. Cho-Jui Hsieh [link]
- **Beijing University of Posts and Telecommunications** Beijing, China
  B.S. *in Department of Computer Science & Technology* *June 2014*
  Ranking 6^{th} **top 2%** among 300 students (Honored 2014 Outstanding Graduate of Beijing)

Project Highlights

- **Building Large-scale Deep Learning Framework for Big Model Area**
  HPC-AI Tech *February 2022 - Now*

  I am the manager of a 10+ member tech team building next-generation open-sourced AI Infra (Github Homepage: https://github.com/hpcaitech). Besides several on-going projects, our open-source software includes:

  1. Colossal-AI [link] is a Unified Deep Learning System for Large-Scale Parallel Training.
  2. Energon-AI [link] is a Large-scale Model Inference System.

- **Building Open-sourced Deep Learning Infrastructures**
  Wechat AI, Tencent *July 2019 - February 2022*

  I was dedicated to solving real production problems in Tencent by proposing innovative system solutions.

  1. I initialized and developed **TurboTransformers** [link], a fast runtime for transformer inference on CPU and GPU.

2. I initialized and developed **PatrickStar** [link], Parallel Training of Large Language Models via a Dynamic Chunk-based Memory Management.
3. Both software is open-sourced on Tencent's official Github and has brought significant cost savings for the company's billion Daily Active User products. I was awarded as the Excellent Contributor for Open-sourced Collaboration of 2021 by Tencent, which is the **highest-valued personal prize** of the company. There are extensive Chinese media reports on my open-source achievements [link], [link].

- **Building Deep Basic Modules for WeChat App**
  Wechat AI, Tencent *July 2019 - March 2021*

  I contributed to a set of basic modules in the WeChat App, including The WeChat Input Engine, the WeChat Open Dialogue Platform, and the WeChat Translation System. WeChat is a super App with over 1 Billion active users per month.

- **Large-scale Deep Learning Training (DL) System for GPU Supercomputer**
  University of California, Davis *September 2017 - August 2018*

  I designed the **RedSync**, a distributed data-parallel Deep Learning training system using gradient pruning and quantization. When scaled up to 128 GPUs, the RedSync brought significant performance improvements to DNNs previously considered hard to scale.

- **High Performance Deep Learning System for the Sunway TaihuLight**
  National Supercomputing Center in Wuxi *April 2016 - August 2019* I

  built a deep learning framework from scratch on the Sunway TaihuLight, which is based on the innovative SW26010 many-core processors and ranked **No.1 on the 47th-50th Top500 Supercomputer lists.**

  1. I designed the **swGEMM** – a GEneral Matrix Multiplication (GEMM) library based on SW26010. Core code is handwritten by the assembly code, reaching 97% of peak performance. Significant speedups (2-10x) were achieved by applying swGEMM instead of default BLAS to deep learning applications.
  2. Designed the **swDNN** – a library that provides APIs for mainstream DL operator (CONV, LSTM, FC, BN, and activations). Regarding the most complicated CONV ops, three parallel schemes were designed for the special SW26010 many-core architecture, i.e. explicit GEMM, implicit GEMM, and Winograd. The computing efficiency of swDNN exceeded cuDNNv7.5 running on Tesla K40.
  3. I designed the **swATOP** – an end-to-end automated framework that optimizes complex parallel DL operator code on SW26010. By reading several lines of DSL statements, swATOP can automatically generate code that exceeds manual optimization performance.
  4. I designed the **swCaffe** – an MPI-based deep learning framework on the Sunway TaihuLight. Synchronization employed an innovative topology-aware MPI Allreduce method which is 10x faster than the default MPI_Allreduce on 1024 nodes.

- **High Performance Scientific Computing Applications**
  Department of Earth System Science, Tsinghua University *February 2014 - March 2016*

  1. I proposed a generalized cache-friendly design based on NVIDIA GPUs and Intel Xeon Phis for complex spatially-variable coefficient (CSVC) stencils. Gained 4x speedup in the seismic imaging software (**GeoEast-Lightning**) used by China National Petroleum Corporation.
  2. I accelerated a series of scientific applications on different HPC platforms, including transient electromagnetic simulation on CPU cluster; remote sensing data analysis with SVM on Intel Xeon Phi; Community Earth System Model (CESM), and crop modeling on Sunway TaihuLight.

First Author Publications [google link]

- **Jiarui Fang**, Yang Yu, Zilin Zhu, Shenggui Li, Yang You, Jie Zhou, **PatrickStar: Parallel Training of Pre-trained Models via Chunk-based Memory Management**, Preprint on arXive. [pdf].
- **Jiarui Fang**, Yang Yu, Chengduo. Zhao, Jie Zhou, **TurboTransformers: An Efficient GPU Serving System For Transformer Models**, Proceedings of the 26th ACM SIGPLAN Symposium on Principles and Practice of Parallel (PPoPP 2021). [pdf] .
- **Jiarui Fang**, Haohuan Fu, Guangwen Yang, Cho-Jui Hsieh, **RedSync : Reducing Synchronization Traffic for Distributed Deep Learning.** Journal of Parallel and Distributed Computing (JPDC), Volume 133, November 2019, Pages 30-39. [arXiv][pdf].
- Wei Gao*, **Jiarui Fang***, Wenlai Zhao, Jinzhe Yang, Long Wang, Lin Gan, Haohuan Fu, Guangwen Yang. **swATOP: Automatically Optimizing Deep Learning Operators on SW26010 Many-Core Processor.** Proceedings of the 48th International Conference on Parallel Processing (ICPP), 2019. (* equal contribution) [pdf] .
- **Jiarui Fang***, and Li, Liandeng* and Fu, Haohuan and Jiang, Jinlei and Zhao, Wenlai and He, Conghui and You, Xin and Yang, Guangwen. **swCaffe: a Parallel Framework for Accelerating Deep Learning Applications on Sunway TaihuLight**, IEEE Cluster (Cluster), Belfast, UK, 2018. [pdf]. (* equal contribution).
- **Jiarui Fang**, Haohuan Fu, Wenlai Zhao, Bingwei Chen, Weijie Zheng, and Guangwen Yang. **swDNN: A library for Accelerating Deep Learning Applications on Sunway Taihulight.** In Parallel and Distributed Processing Symposium (IPDPS), 2017 IEEE International, pages 615–624. IEEE, 2017. [pdf]
- **Jiarui Fang**,, Haohuan Fu, Guangwen Yang. **Cache-friendly Design for Complex Spatially-variable Coefficient Stencils on Many-core Architectures**. IEEE 23rd International Conference on High Performance Computing, Data, and Analytics (HiPC),p222-p231, Hyderabad, India, 2016. [pdf]
- **Jiarui Fang**, Haohuan Fu, He Zhang, Wei Wu, Nanxun Dai, Lin Gan, Guangwen Yang. **Optimizing Complex Spatially-Variant Coefficient Stencils for Seismic Modeling on GPU**. IEEE 21st International Conference on Parallel and Distributed Systems (ICPADS), p641-p648 Melbourne, Australia, 2015. [pdf]
- **Jiarui Fang**, Haohuan Fu, Guangwen Yang, Wei Wu, Nanxun Dai. **GPU-based explicit time evolution method.** The 84th Society of Exploration Geophysicists Technical Program Expanded Abstracts (SEG), p3549-p3553, New Orleans, USA, 2015 [pdf]

Skills

- **Good at English:** GRE 322 (Verbal 153, October 2012)
- **Programing Language:** C/C++, CUDA, Python
- **Technical Skills:** Computer Architecture, Parallel Computing, Software Performance Tuning and Optimization, Deep Learning, Numerical Computing.

Academia Service

- I serve as the reviewer of journals include ACM Transactions on Architecture and Code Optimization (TACO), Parallel Computing (PARCO), Transactions on Cloud Computing (TOCC), Transactions on Parallel and Distributed Systems (TPDS).

References

- **Jie Zhou**
  Director of the Pattern Recognition Center, WeChat AI.
  Email:withtomzhou@tencent.com

- **Guangwen Yang**
  Professor in Department of Computer Science, Tsinghua University,
  Director of the National Supercomputing Center in Wuxi.
  Email:ygw@tsinghua.edu.cn
- **Haohuan Fu**
  Professor in Department of Earth Science, Tsinghua University,
  Deputy Director of the National Supercomputing Center in Wuxi.
  Email:haohuan@tsinghua.edu.cn
- **Cho-Jui Hsieh**
  Assistant Professor in Department of Computer Science, University of California, Los Angeles.
  Email:chohsieh@cs.ucla.edu