Jiarui Fang (方佳瑞)

☎ +(86) TBD ✉ fangjiarui123@gmail.com
https://github.com/feifeibear (600+ followers)
Update Date: June 15, 2024

I have extensive experience in research and engineering, specifically in High-Performance Computing (HPC) and Machine Learning Systems (MLSys). I have been the contributor to several popular open-source MLSys software projects with over **40k** stars on GitHub. I am the first author of several publications in top-tier computer science conferences and journals.

Work Experience

- **Tencent Clould** Beijing, China
  Principal Software Engineer, Tech Leader *November 2023 - Now*
- **LightYearAI-Meituan (startup)** Beijing, China
  Technical Partner *April 2023 - November 2023*
- **HPC-AI Technology (startup)** Beijing, China
  Co-founder, Chief Technology Officer *February 2022 - March 2023*
- **Tencent WeChat AI** Beijing, China
  Senior Software Engineer *July 2019 - February 2022*

Education

- **Tsinghua University** Beijing, China
  Ph.D. *in Department of Computer Science & Technology* *September 2014 - July 2019*
  Advisor: Prof. Guangwen Yang, Co-advisor: Prof. Haohuan Fu
  Dissertation: Parallel Deep Learning Training System on Sunway TaihuLight [pdf]
- **University of California, Davis** Davis, CA, USA
  Visiting Scholar *in Department of Computer Science* *August 2017 - August 2018*
  Advisor: Associate Prof. Cho-Jui Hsieh [link]
- **Beijing University of Posts and Telecommunications** Beijing, China
  B.Eng. *in Department of Computer Science & Technology* *September 2010 - June 2014*
  Ranking 6th **top 2%** among 300 students (Honored 2014 Outstanding Graduate of Beijing)

Project Highlights

- **Building Large Language/Vision Model Systems on Public Cloud**
  Tencent *November 2023 - Now*

  Tech Leader of a team for developing Large Language/Vision Model training and inference software on heterogeneous hardware platforms, including GPU and NPUs. Contributed some key technologies to the community such as:

  1. USP, a unified sequence parallelism approach, is utilized by companies like Huawei, Alibaba and Sensetime to train long sequence LLMs and Diffusion Models.
  2. PipeFusion, an efficient parallel inference of diffusion transformer models, is used for the deployment of large visual models like Sora.

- **Data Curation for Large Language Model (LLM) Pre-training**
  LightYear-AI *April 2023 - August 2023*

  As the first technical member at LightYear AI, a unicorn AGI startup company with over 230 million USD in funding, I led a team of more than 20 professionals. Our primary responsibilities encompassed data collection, curation, and quality evaluation for LLM pre-training. This company was subsequently acquired by the renowned Meituan Inc., just three months post-establishment.

- **Building Open-sourced LLM System**
  HPC-AI Technology *February 2022 - March 2023*

  Serving as the CTO at HPC-AI Tech, a startup that successfully raised over 9 million USD in angel round financing, I am entrusted with the leadership of a dynamic team of 10+ members. We are at the forefront of developing next-generation open-source AI infrastructures (GitHub homepage: https://github.com/hpcaitech). Among several innovative projects, I have been instrumental in leading the development of the following open-source software:

  1. **Colossal-AI** [link] is a unified deep learning system for large-scale parallel training in the era of big models. Under my leadership, the project was widely adopted and garnered **18k** GitHub stars within 12 months, up from 0.8k stars.
  2. **Energon-AI** [link] is a large-scale language model inference system.
  3. **FastFold** [link] is a training and inference system for AI-based protein structure prediction on GPU clusters.

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

  At Tencent, I was dedicated to solving real production AI problems by proposing innovative HPC system solutions. My accomplishments include:

  1. I initiated and developed **TurboTransformers** [link], a fast runtime for transformer inference on CPU and GPU.
  2. I initiated and developed **PatrickStar** [link], a large language model training framework featuring dynamic chunk-based memory management. This was the first solution that was able to train GPT3 on 4 NVIDIA SuperPod nodes with 32 GPUs.
  3. Both software packages are open-sourced on Tencent's official GitHub and have brought significant cost savings for the company's billion Daily Active User products.

  My achievements were recognized by Tencent with the **highest-valued personal prize**, the Excellent Contributor for Open-sourced Collaboration of 2021 award. Chinese media have reported extensively on my open-source contributions, which can be found at [link] and [link].

- **Building 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 Method Engine (C++), the WeChat Open Dialogue Platform (C++), and the WeChat Translation System (PyTorch). WeChat is a super App with over 1 Billion active users per month.

- **Building Deep Learning Training 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 on Piz Daint Supercomputer (the No.5 fastest supercomputer at that time), the RedSync brought significant performance improvements to DNNs previously considered hard to scale.

- **Building Deep Learning Training System for the Sunway TaihuLight Supercomputer**
  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. I designed the **swDNN** – a library that provides APIs for mainstream deep learning operators (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 deep learning operator implementation 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.

- **Optimizing for Parallel Software of Scientific Computing Applications**
  Tsinghua University *February 2014 - March 2016*

  1. I proposed a generalized cache-friendly design based on NVIDIA GPUs and Intel Xeon Phis (CUDA/C++) for complex spatially-variable coefficient 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.

Selected Publications [google scholar link]

1. Xuanlei Zhao, Shenggan Cheng, Guangyang Lu, **Jiarui Fang**, Haotian Zhou, Bin Jia, Ziming Liu, Yang You **AutoChunk: Automated Activation Chunk for Memory-Efficient Long Sequence Inference.** The 2024 International Conference on Learning Representations (ICLR 2024)

2. Shenggan Cheng, Xuanlei Zhao, Guangyang Lu, **Jiarui Fang**, Zhongming Yu, Tian Zheng, Ruidong Wu, Xiwen Zhang, Jian Peng, Yang You **FastFold: reducing AlphaFold training time from 11 days to 67 hours.** Proceedings of the 29th ACM SIGPLAN Annual Symposium on Principles and Practice of Parallel Programming (PPoPP 2023)

3. Shenggui Li, Hongxin Liu, Zhengda Bian, **Jiarui Fang**, Haichen Huang, Yuliang Liu, Boxiang Wang, Yang You. **Colossal-AI: A Unified Deep Learning System For Large-Scale Parallel Training.** Proceedings of the 52nd International Conference on Parallel Processing (ICPP 2023). 2023: 766-775.

4. **Jiarui Fang**, Zilin Zhu, Shenggui Li, Hui Su, Yang Yu, Jie Zhou, Yang You. **Parallel Training of Pre-trained Models via Chunk-based Dynamic Memory Management**, in IEEE Transactions on Parallel and Distributed Systems (TPDS), 2022, 34(1): 304-315. [pdf].

5. Hui Su, Weiwei Shi, Xiaoyu Shen, Zhou Xiao, Tuo Ji, **Jiarui Fang**, Jie Zhou. **RoCBert: Robust Chinese Bert with Multimodal Contrastive Pretraining.** In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2021) (pp. 921-931).

6. **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] .

7. Liandeng Li, **Jiarui Fang**, Jinlei Jiang, Lin Gan, Weijie Zheng, Haohuan Fu, Guangwen Yang. **Efficient AES implementation on Sunway TaihuLight supercomputer: A systematic approach.** Journal of Parallel and Distributed Computing (JPDC), 2020, 138: 178-189.

8. **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].

9. 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] .

10. Li, Liandeng* and **Jiarui, Fang*** 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 Belfast, UK, (Cluster 2018), [pdf]. (* equal contribution).

11. Wenlai Zhao, Haohuan Fu, **Jiarui Fang**, Weijie Zheng, Lin Gan, Guangwen Yang. **Optimizing Convolutional Neural Networks on the Sunway Taihulight Supercomputer.** ACM Transactions on Architecture and Code Optimization (TACO), 2018, 15(1): 1-26.

12. **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), 2017 IEEE International, pages 615–624. IEEE, 2017. [pdf]

13. **Jiarui Fang**, Haohuan Fang, 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 16'),p222-p231, Hyderabad, India, 2016. [pdf]
14. **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 15'), p641-p648 Melbourne, Australia, 2015. [pdf]

Selected Preprints

1. Jiannan Wang*, **Jiarui Fang***, Aoyu Li, PengCheng Yang **PipeFusion: Displaced Patch Pipeline Parallelism for Inference of Diffusion Transformer Models.** (* equal contribution) arXiv preprint arXiv:2405.14430

2. **Jiarui Fang**, Shangchun Zhao **USP: A Unified Sequence Parallelism Approach for Long Context Generative AI.** preprint arXiv:2405.07719, 2023.

3. Haichen Huang, **Jiarui Fang**, Hongxin Liu, Shenggui Li, Yang You **Elixir: Train a Large Language Model on a Small GPU Cluster.** preprint arXiv:2212.05339, 2023.

Patents

I authored 21 Chinese patents whose public numbers are as follows (* the first inventor):

- **HPC-AI Technoloy:** CN114860445A*, CN115204369A*, CN115687232A, CN115860101A, CN115719619A, CN116050512A, CN115423092A, CN115061804A, CN115374909A, CN115061804A, CN114816801A, CN115480702A, CN115455150A.
- **Tencent:** CN114282665A*, CN114330700A*, CN111898698A*, CN111708641A*, CN111475775A*, CN111488177A*, CN114444476A.
- **NSCCWX:** CN110929850A.

Skills

- **Good at English:** CET-6 (591) , The Public English Test System Level 5 (WSK-PETS5) Certification
- **Programming Language:** C/C++, CUDA, Python

Academia Service

I serve as a reviewer of the following journals:

- Transactions on Parallel and Distributed Systems (TPDS)
- Journal of Parallel and Distributed Computing (JPDC)
- Journal of Supercomputing
- ACM Transactions on Architecture and Code Optimization (TACO)
- Parallel Computing (PARCO)
- Transactions on Cloud Computing (TCC)
- Cluster Computing
- Pattern Recognition
- IEEE Access

References

- **Huiwen Wang**
  Co-founder of Meituan Inc & Co-founder of Lightyear AI.
  Email:wanghuiwen1978@gmail.com
- **Jie Zhou**
  Director of the Pattern Recognition Center, WeChat AI, Tencent.
  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**
  Associate Professor in Department of Computer Science, University of California, Los Angeles.
  Email:chohsieh@cs.ucla.edu