Jiarui Fang (方佳瑞)

+(86) 13717819702 fangjiarui123@gmail.com
https://github.com/feifeibear (290+ followers)
Update Date: January 17, 2023

Extensive research and engineering experience in high-performance advanced computing and artificial intelligence technology. The manager of an international technical team with 20+ members. The main contributor to popular open-source software that has earned over **10k stars** on GitHub. The first author of publications on top-tier computer science conferences and journals.

Work Experience

- **HPC-AI Technology** Beijing, China
  Co-founder, Chief Technology Officer *February 2022 - Now*
- **WeChat AI, Tencent** Beijing, China
  Senior Research Scientist *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* *April 2016 - August 2017*
  Mentor: Prof. Haohuan Fu

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 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 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 Technology *February 2022 - Now*

  I am the manager of a 20+ member tech team building next-generation open-sourced AI Infra (Github Homepage: https://github.com/hpcaitech). Besides several ongoing projects, our open-source software includes:

  1. **Colossal-AI** [link] is a Unified Deep Learning System for Large-Scale Parallel Training in Big Model Era.
  2. **Energon-AI** [link] is a Large-scale Model Inference System.
  3. **FastFold** [link] is Model Training and Inference system for Protein Structure Prediction on GPU Clusters.

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

  I was dedicated to solving real production AI problems in Tencent by proposing innovative HPC 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], a large language Model training framework features with 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 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 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.

- **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 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.

- **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. I 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.

Publications [google scholar link]

1. **Jiarui Fang**, Yang Yu, Zilin Zhu, Shenggui Li, Yang You, Jie Zhou, **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].

2. 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).

3. **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] .

4. 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.

5. **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].

6. 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] .

7. 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).

8. Weijia Li*, Conghui He*, **Jiarui Fang**, Juepeng Zheng, Haohuan Fu, Le Yu. **Semantic segmentation-based building footprint extraction using very high-resolution satellite images and multi-source GIS data.** Remote Sensing, 2019, 11(4): 403. (* equal contribution)

9. Xiao Huang, Chaoqing Yu, **Jiarui Fang**, Guorui Huang, Shaoqiang Ni, Jim Hall, Conrad Zorn, Xiaomeng Huang, Wenyuan Zhang. **A dynamic agricultural prediction system for large-scale drought assessment on the Sunway TaihuLight supercomputer.** Computers and electronics in agriculture, 2018, 154: 400-410.

10. 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.

11. Weijia Li*, Conghui He*, **Jiarui Fang**, Haohuan Fu. **Semantic segmentation based building extraction method using multi-source gis map datasets and satellite imagery.**Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2018: 238-241. (* equal contribution)

12. Liandeng Li, **Jiarui Fang**, Jinlei Jiang, Lin Gan, Weijie Zheng, Haohuan Fu, Guangwen Yang. **SW-AES: accelerating AES algorithm on the sunway taihulight.** 2017 IEEE International Symposium on Parallel and Distributed Processing with Applications and 2017 IEEE International Conference on Ubiquitous Computing and Communications (ISPA/IUCC 2017). IEEE, 2017: 1204-1211.

13. Weijia Li, Haohuan Fu, Yang You, Le Yu, **Jiarui Fang**. **Parallel multiclass support vector machine for remote sensing data classification on multicore and many-core architectures.** IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2017, 10(10): 4387-4398

14. **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]

15. Haohuan Fu, Junfeng Liao, Wei Xue, et al. **Refactoring and optimizing the community atmosphere model (CAM) on the sunway taihulight supercomputer.**: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC 2016). IEEE, 2016: 969-980.

16. **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 2016),p222-p231, Hyderabad, India, 2016. [pdf]

17. **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 2015), p641-p648 Melbourne, Australia, 2015. [pdf]

18. **Jiarui Fang**, Haohuan Fu, Guangwen Yang, Wei Wu, Nanxun Dai. **GPU-based explicit time evolution method.** 2015 SEG Annual Meeting, p3549-p3553, New Orleans, USA, (SEG 2015) [pdf]

19. Haohuan Fu, Yingqiao Wang, Evan Schankee Um, **Jiarui Fang**, Tengpeng Wei, Xiaomeng Huang, Guangwen Yang. **A parallel finite-element time-domain method for transient electromagnetic simulation.** Geophysics, 2015, 80(4): E213-E224.

Preprints

1. Haichen Huang, **Jiarui Fang**, Hongxin Liu, Shenggui Li, Yang You. **Elixir: Train a Large Language Model on a Small GPU Cluster.** arXiv preprint arXiv:2212.05339, 2022.

2. Jiangsu Du, Ziming Liu, **Jiarui Fang**, Shenggui Li, Yongbin Li, Yutong Lu, Yang You. **EnergonAI: An Inference System for 10-100 Billion Parameter Transformer Models.** arXiv preprint arXiv:2209.02341, 2022.

3. **Jiarui Fang**, Geng Zhang, Jiatong Han, Shenggui Li, Zhengda Bian, Yongbin Li, Jin Liu, Yang You. **A Frequency-aware Software Cache for Large Recommendation System Embeddings.** arXiv preprint arXiv:2208.05321, 2022.

4. Shenggui Li, **Jiarui Fang**, Zhengda Bian, Hongxin Liu, Yuliang Liu, Haichen Huang, Boxiang Wang, Yang You. **Colossal-AI: A Unified Deep Learning System For Large-Scale Parallel Training.** arXiv preprint arXiv:2110.14883, 2021.

Patents

I authored 14 Chinese patents whose public numbers are as follows (* the first inventor):

- **HPC-AI Technoloy:** CN115423092A, CN115374909A, CN115204369A, CN115061804A, CN114860445A*, CN114816801A.
- **Tencent:** CN114444476A, CN114282665A*, CN114330700A*, CN111898698A*, CN111708641A*, CN111475775A*, CN111488177A*.
- **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)
- ACM Transactions on Architecture and Code Optimization (TACO)
- Parallel Computing (PARCO)
- Transactions on Cloud Computing (TCC)
- IEEE Access
- Cluster Computing
- Pattern Recognition

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**
  Associate Professor in Department of Computer Science, University of California, Los Angeles.
  Email:chohsieh@cs.ucla.edu