Shenggan Cheng

+(86) 13636684726 +(65) 98854484

shenggan@comp.nus.edu.sg shenggan.c@outlook.com

EDUCATION

Shanghai Jiao Tong University (B.E. in Computer Science) *2015.9 – 2019.6*

National University of Singapore (Ph.D. in Computer Science) *2022.01 – now*

WORK & RESEARCH EXPERIENCE

National University of Singapore, HPC-AI Lab (Prof. Yang You) *2022.01 – now*

FastFold: Optimizing AlphaFold Training and Inference on GPU Clusters (PPoPP 2024)

https://github.com/hpcaitech/FastFold

FastFold is the first performance optimization method for the training and inference of protein structure prediction models. It significantly reduces the time and economic costs of AlphaFold model training and inference by applying large model training techniques such as parallelism strategies and communication optimization.

Hanayo: Wave-like Pipeline Parallelism for LLM Training (SC 2023)

Hanayo proposed a wave-like pipeline parallelism strategy, alongside a high-performance pipeline execution runtime. Hanayo mitigates the issues of pipeline bubbles and excessive memory consumption prevalent in existing schemes, without resorting to model duplicates as in Chimera.

Alibaba Cloud, PAI team, Alibaba Innovative Research *2023.02 – 2024.08*

Automated Communication Optimization and Scheduling (ASPLOS 2025)

Concerto introduced a compiler framework designed to address computation-communication overlapping challenges by automatically optimizing and scheduling communication. We formulate the scheduling problem as a resource-constrained project scheduling problem and use off-the-shelf solver to get the near-optimal scheduling. And use auto-decomposition to create overlap opportunity for critical (synchronous) communication.

Shanghai Jiao Tong University, Center for HPC, Research Engineer *2019.06 – 2021.04*

Parallel Cosmology N-Body Simulation for HPC Cluster (CCGrid 2020)

Design and develop a cosmology N-body code with Tsung-Dao Lee Institute, SJTU and Xiamen University, which conducted Cosmo-pi simulation on a 2 PFlops Intel cluster with 20,480 cores with highly performance and scalability.

PUBLICATION

Conference Paper

Shenggan Cheng*, Shengjie Lin*, Lansong Diao, Hao Wu, Siyu Wang, Chang Si, Ziming Liu, Xuanlei Zhao, Jiangsu Du, Wei Lin, Yang You. *"Concerto: Automatic Communication Optimization and Scheduling for Large-Scale Deep Learning". ASPLOS 2025 (CCF A)*

Shenggan Cheng, Xuanlei Zhao, Guangyang Lu, Jiarui Fang, Tian Zheng, Ruidong Wu, Xiwen Zhang, Jian Peng, Yang You. *"FastFold: Optimizing AlphaFold Training and Inference on GPU Clusters". PPoPP 2024 (CCF A)*

Ziming Liu*, **Shenggan Cheng***, Haotian Zhou, Yang You. *"Hanayo: Harnessing Wave-like Pipeline Parallelism for Enhanced Large Model Training Efficiency". SC 2023 (CCF A)*

Shenggan Cheng*, Hao-Ran Yu*, Derek Inman, Qiucheng Liao, Qiaoya Wu, James Lin. *"CUBE--Towards an Optimal Scaling of Cosmological N-body Simulations". CCGRID 2020 (CCF C)*

*Xuanlei Zhao, **Shenggan Cheng**, Chang Chen, Zangwei Zheng, Ziming Liu, Zheming Yang, Yang You "DSP: Dynamic Sequence Parallelism for Multi-Dimensional Transformers". ICML 2025 (CCF A)*

*Yong Liu, Di Fu, **Shenggan Cheng**, Zirui Zhu, Yang Luo, Minhao Cheng, Cho-Jui Hsieh, Yang You "SeedLoRA: A Fusion Approach to Efficient LLM Fine-Tuning". ICML 2025 (CCF A)*

Xuanlei Zhao, **Shenggan Cheng**, Guangyang Lu, Haotian Zhou, Bin Jia, Yang You. *"AutoChunk: Automated Activation Chunk for Memory-Efficient Long Sequence Inference ". ICLR 2024*

Li, Binrui, **Shenggan Cheng**, and James Lin. "*tcFFT: Accelerating Half-Precision FFT through Tensor Cores." CLUSTER 2021 (CCF B)*

Jiangsu Du, Jinhui Wei, Jiazhi Jiang, **Shenggan Cheng**, Dan Huang, Zhiguang Chen, Yutong Lu. *"Liger: Interleaving Intra- and Inter-Operator Parallelism for Distributed Large Model Inference". PPoPP 2024 (CCF A)*

Xuanlei Zhao, Bin Jia, Haotian Zhou, Ziming Liu, **Shenggan Cheng**, Yang You. *"HeteGen: Efficient Heterogeneous Parallel Inference for Large Language Models on Resource-Constrained Devices". MLSys 2024*

Kunyuan Du, Ya Zhang, Haibing Guan, Qi Tian, **Shenggan Cheng**, James Lin. *"FTL: A universal framework for training low-bit DNNs via Feature Transfer". ECCV 2020 (CCF B)*

Preprint

Shenggan Cheng, Ziming Liu, Jiangsu Du, Yang You. *"ATP: Adaptive Tensor Parallelism for Foundation Models". arXiv 2023*

Ziming Liu, Shaoyu Wang, **Shenggan Cheng**, Zhongkai Zhao, Yang Bai, Xuanlei Zhao, James Demmel, Yang You. *"WallFacer: Guiding Transformer Model Training Out of the Long-Context Dark Forest with N-body Problem". arXiv 2024*

Shenggan Cheng, Yuanxin Wei, Lansong Diao, Yong Liu, Bujiao Chen, Lianghua Huang, Yu Liu, Wenyuan Yu, Jiangsu Du, Wei Lin, Yang You. *"SRDiffusion: Accelerate Video Diffusion Inference via Sketching-Rendering Cooperation". arXiv 2025*

Journal Article

Zixuan Huang, Junming Fan, **Shenggan Cheng**, Shuai Yi, Xiaogang Wang, Hongsheng Li. "HMS-Net: Hierarchical Multi-Scale Sparsity-Invariant Network for Sparse Depth Completion". TIP 2019 *(CCF A)*

HONORS

ASC18 Student Supercomputer Competition	*First Prize (4th Worldwide)*	*2018.3*
The 5th Student RDMA Programming Competition	*First Prize*	*2017.10*
The Interdisciplinary Contest in Modeling	*Meritorious Winner*	*2017.2*