Shenggan Cheng

+(86) 13636684726 +(65) 98854484

shenggan@comp.nus.edu.sg

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

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.

- *Performance optimization based on the AlphaFold-specific characteristic. Combined with kernel fusion, our kernel implementation of FastFold achieves significant speedups.*
- *We propose Dynamic Axial Parallelism and Duality Async Operation which has a lower communication overhead than other model parallelism methods.*

Alibaba Cloud, PAI team, Alibaba Innovative Research *2023.02 – 2024.04*

EasyDist: Automated Parallelization System and Infrastructure

https://github.com/alibaba/easydist

EasyDist is an automated parallelization system and infrastructure designed for multiple ecosystems, offering the following key features:

- ***Usability**: parallelizing training or inference code to a larger scale becomes effortless with just a single line of change.*
- ***Ecological Compatibility**: EasyDist serves as a centralized source of truth for SPMD rules at the operator-level for various machine learning frameworks*
- *EasyDist decouples auto-parallel algorithms from specific machine learning frameworks and IRs.*

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

Parallel Cosmology N-Body Simulation for HPC Cluster

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.

SenseTime Research Department, Research Intern *2018.3 – 2019.4*

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*

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*

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*

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

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*

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*

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*

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*

Preprint

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

Xuanlei Zhao, **Shenggan Cheng**, Zangwei Zheng, Zheming Yang, Ziming Liu, Yang You*. "DSP: Dynamic Sequence Parallelism for Multi-Dimensional Transformers". arXiv 2024*

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*

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