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) *2021.10 – now*

FastFold: Reducing AlphaFold Training Time from 11 Days to 67 Hours

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. (DAP is used by BladeDISC example for AlphaFold model)*

ATP: Adaptive Tensor Parallelism for Foundation Models

ATP is an adaptive tensor parallelism framework for foundation models, which can automatically select the optimal parallel strategy on different interconnections.

- *We propose column- and row-first tensor parallelism based on 2D device meshes and construct a search space. Combined with the hierarchical communication matrix, ATP can identify the optimal strategy in the search space.*
- *The theoretical model of ATP shows that the communication overhead of ATP decreases with scaling, indicating a qualitative leap forward.*

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

AI Platform and HPC Container Platform

The chief operator of SJTU AI Platform which consisted by 8 NVIDIA DGX-2 Supercomputer (128 Volta GPU with highly speed connection). Participated in all the works from platform design, construction to formal operation.

- *System Integrated: Parallel Filesystem, Job Scheduler and Container using Automated Deployment Tools.*
- *Propose and Implement Unprivileged User Container Build Solution for HPC System.*

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. Cosmo-pi simulation contains 4.4 trillion particles, break the world record of N-Body simulation.

- *Designed fixed-point compression to obtain the lowest memory consumption possible.*
- *High performance MPI-OpenMP hybrid Implement with compression-aware optimization.*

SenseTime Research Department, Research Intern *2018.3 – 2019.4*

Prediction of Pedestrians and Vehicle Trajectory

This R&D project is conducted by SenseTime which aim to design and develop a module to predict the trajectory of pedestrians and Vehicle. The prediction is based on the results of scene recovery and the results of prediction will be the input of Path Planning and Decision Module.

HONORS

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

PUBLICATION

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". arXiv 2022*

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

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*

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

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*

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*