# Xuanlei, ZHAO

xuanlei [at] comp.nus.edu.sg | +86-13817229553
Homepage | Github | Singapore

## EDUCATIONAL BACKGROUND

**National University of Singapore** *2024.01 - Present*
- Ph.D. in Computer Science

**National University of Singapore** *2022.08 - 2023.12*
- M.S. in Computer Science

**Huazhong University of Science and Technology** *2018.09 - 2022.06*
- B.Eng. in Computer Science & Electronic Information

## RESEARCH EXPERIENCE

### OpenDiT: An Easy, Fast and Memory-Efficient System for DiT Training and Inference

*National University of Singapore / Advised by Yang You* *2024.02 - Present*

*Objective: The first and best open-source project to accelerate distributed training and inference of Diffusion Transformer (foundation model of Sora). Gain over 600 stars on Github within 4 days.*

- Propose FastSeq, a novel sequence parallelism for intra-node and long sequence parallel, reducing 50% communication time compared with sota methods by reducing communication cost and introducing efficient and asynchronized operations.
- Combine an initiative ema sharding strategy with ZeRO and mixed precision to reduce memory cost.
- Introduce a fused adaLN kernel to reduce I/O overhead and enhance computational efficiency, incorporating kernels including FlashAttention, Fused Layernorm, and FusedAdam.
- Achieve 80% speedup and 50% reduction in memory costs for DiT training.

### AutoChunk: Automated Activation Chunk for Memory-Efficient Deep Learning Inference

*National University of Singapore / Advised by Yang You* *2022.11 - 2023.06*

*Objective: To improve the efficiency of inference for models where activation takes up most memory, we proposed AutoChunk, a compiler system to reduce memory cost and speed up inference by automatically decomposing activation with minimal performance loss.*

- Introduce an innovative method for enabling automated activation chunk for general models during inference to reduce activation memory effectively.
- Design a bottom-up scheduler to search the best decomposition strategy that reduces memory usage with least performance lost based on the observation of the uneven distribution of memory cost.
- Reduce 80% of activation memory while maintaining speed loss within 10% and reduce up to 99.9% if not considering speed.

### HeteGen: Efficient Heterogeneous Generative Inference of LLMs on Low-Resource Devices

*National University of Singapore / Advised by Yang You* *2023.04 - 2023.10*

*Objective: To improve the efficiency of offload-based inference for LLMs on low-resource devices, we proposed HeteGen, a heterogeneous parallel system to distribute workload between CPU and GPU to maximize computation efficiency and alleviate parameter I/O bottleneck.*

- Derive a general theoretical formula to balance computation on CPU and GPU for offload-based inference and apply it to large language models.
- Design an asynchronous scheduling algorithm parallel heterogeneous computing, which effectively reduces communication overhead and maximizes the overlap between communication and computation.
- Improve latency by 317% at most compared with the current state-of-the-art methods.

# Xuanlei, ZHAO

xuanlei [at] comp.nus.edu.sg | +86-13817229553
Homepage | Github | Singapore

**FastFold: Optimizing AlphaFold Training and Inference on GPU Clusters**

*National University of Singapore / Advised by Yang You* *2022.7 - 2023.2*

*Objective: To reduce the cost of training and inference of AlphaFold, we proposed FastFold, the first method that significantly improves its performance by novel parallelism method, specialized kernels, and memory optimizations.*

- Design shared memory and heterogenous computing to further optimize memory by 30%, making it able to cover the length of 99.9% protein on one GPU.
- Help to propose Dynamic Axial Parallelism with Duality Async Operation, reducing communication overhead and improves overall performance by 24.14% compared with tensor parallelism.
- Help to design optimized and fused kernels based on the AlphaFold-specific characteristic to achieve up to 3.32x speedup.

## WORK EXPERIENCE

**HPC-AI TECH Inc. (Colossal-AI)** *Singapore*

*MLSys Research Intern / Advised by Jiarui Fang and Yang You* *2022.5 - 2023.12*

- Contribute 35k lines of code as a core contributor (rank 5th) and help it gain over 35k stars on Github.
- Propose BalanceMoE, an expert parallelism focus on reducing the cost of unbalance computation and communication from three levels, which improves throughput by 42% compared with sota.
- Utilize AutoChunk to support real applications such as AlphaFold, Stable Diffusion and ChatGPT.
- Actively participate in the development of various parallelism strategies including hybrid parallelism, ZeRO, Offload, auto parallelism and various efficient kernels.

## PUBLICATIONS

- **Xuanlei Zhao**, Shenggan Cheng, Guangyang Lu, Haotian Zhou, Bin Jia. *AutoChunk: Automated Activation Chunk for Memory-Efficient Deep Learning Inference. ICLR 2024*
- Shenggan Cheng, **Xuanlei Zhao**, Guangyang Lu, Jiarui Fang, Zhongming Yu, Tian Zheng, Ruidong Wu, Xiwen Zhang, Jian Peng, and Yang You. *FastFold: Reducing AlphaFold Training Time from 11 Days to 67 Hours. PPoPP 2024*
- **Xuanlei Zhao***, Bin Jia*, Haotian Zhou*, Ziming Liu, Shenggan Cheng, and Yang You. *HeteGen: Efficient Heterogeneous Parallel Inference for Large Language Models on Resource-Constrained Devices. MLSys 2024*