An Optimization Technique for Hiding Communication Costs in 3D Parallel Training of DL
(IEEE/ACM CCGrid2025 [59])

Ryubu Hosoki [†1], Kento Sato [†2], Toshio Endo [†1], Julien Bigot [†3], Edouard Audit [†3]
†1: Institute of Science Tokyo, †2: RIKEN R-CCS, †3: CEA

Best Paper Award

- **Background**
  - DNN models have grown rapidly in accuracy, but this progress has come with a training cost (e.g., 100Bs-Ts params)
  - Training large models takes tremendous time and memory capacity → Parallel training, but challenging in decomposition
  - Auto-parallelization (e.g., Alpa) finds the optimal balance of DP/TP/PP without expertise for parallel training in HPC systems
  - XLA is a domain-specific compiler, XLA: high-level computation graph (e.g., JAX, PyTorch) → Optimized machine code
  - An XLA compiler in Alpa produces inefficient all-reduce stages in PP backward computation [**Fig.1**]
- **Approach: Comm-Shift Optimization at the level of an XLA compiler used in Alpa** *XLA: Accelerated Linear Algebra
  - We analyze computation graph and shift gradient-averaging communication from backward to parameter update [**Fig.2**]
    → eliminate synchronization time from one pipeline stage from another → Reduce overall training times
- **Results**
  - Comm-Shift improves the training performance across various models up to 27% at maximum (GPT-J-6B) [**Table 1**]
  - Improvement becomes more significant with more communication in larger models

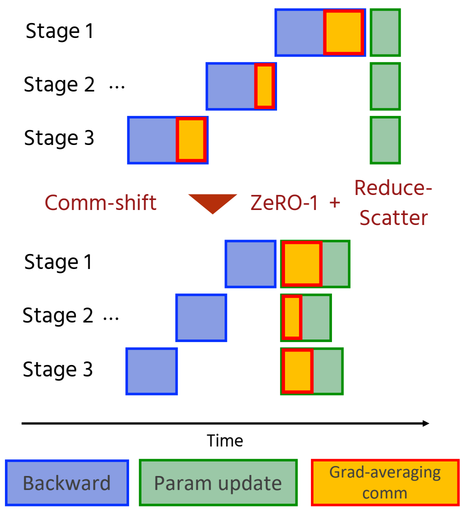

Fig.1

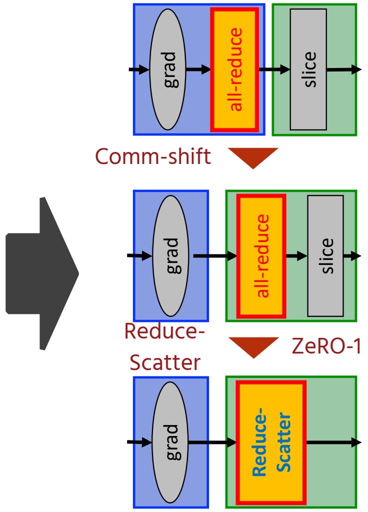

Fig.2

Table 1: Experimental results on TSUBAME4.0 (NVIDIA H100 SXM5 /InfiniBand NDR200 x4 (Fat Tree))

Task	Model	# of GPUs	Strategy	Throughput w/o comm-shift opt.	Throughput w/ comm-shift opt.	Speedup
NLP	GPT-2 Small	16	[(8x1), (8x1)]	2019190 token/s	2073480 token/s	+2.69%
	GPT-2 Large	16	[(8x1), (8x1)]	403880 token/s	419156 token/s	+3.78%
	GPT-2 XL	32	[(16x1), (16x1)]	477236 token/s	508616 token/s	+6.58%
	GPT-J-6B	32	[(8x1), (4x2), (4x2), (4x2)]	143158 token/s	181812 token/s	+27.00%
	Mamba-1.4B	16	[(2x1), (2x1), (2x1), (2x1), (2x1), (2x1), (2x1), (2x1)]	29786 token/s	30913 token/s	+3.79%
Image Classification	ViT-base	8	[(2x1), (2x1), (2x1), (2x1)]	1192 image/s	1306 image/s	+9.55%
	SwinV2-L	16	[(4x1), (2x1), (4x1), (4x1), (2x1)]	1827 image/s	1893 image/s	+3.60%
	CoAtNet-7	16	[(8x1), (8x1)]	384 image/s	444 image/s	+15.36%

[59] Ryubu Hosoki, Kento Sato, Toshio Endo, Julien Bigot, Edouard Audit, "An Optimization Technique for Hiding Communication Costs in 3D Parallelism", In the proceedings of The IEEE International Symposium on Cluster, Cloud, and Internet Computing (CCGrid 2025), May, 2025