NEURAL INFORMATION
PROCESSING SYSTEMS

SDP4Bit: Toward 4-bit Communication Quantization in Sharded Data Parallelism for LLM Training

Jinda Jia , Cong Xie*, Hanlin Lu, Daoce Wang, Hao Feng, Chenmgming Zhang, Baixi Sun, Haibin Lin, Zhi Zhang, Xin Liu, Dingwen Tao†*

INDIANA UNIVERSITY

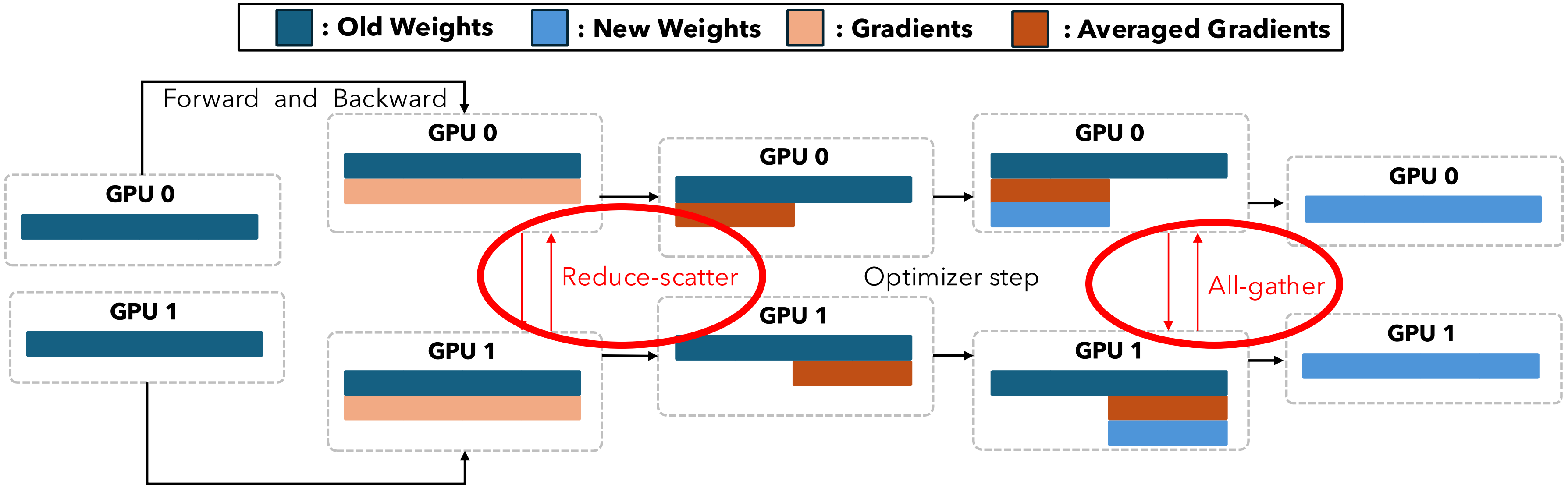
ByteDance

* Equaly contributed first authors

† *Correspoding author*

Communication Overhead is Large During Training

Communication is slow, especially inter-node communication.



Communication pattern for Shared Data Parallelism

What SDP4Bit Can Achieve?

NEURAL INFORMATION PROCESSING SYSTEMS

Weight Communication:

16 bits → 4 bits

Gradient Communication :

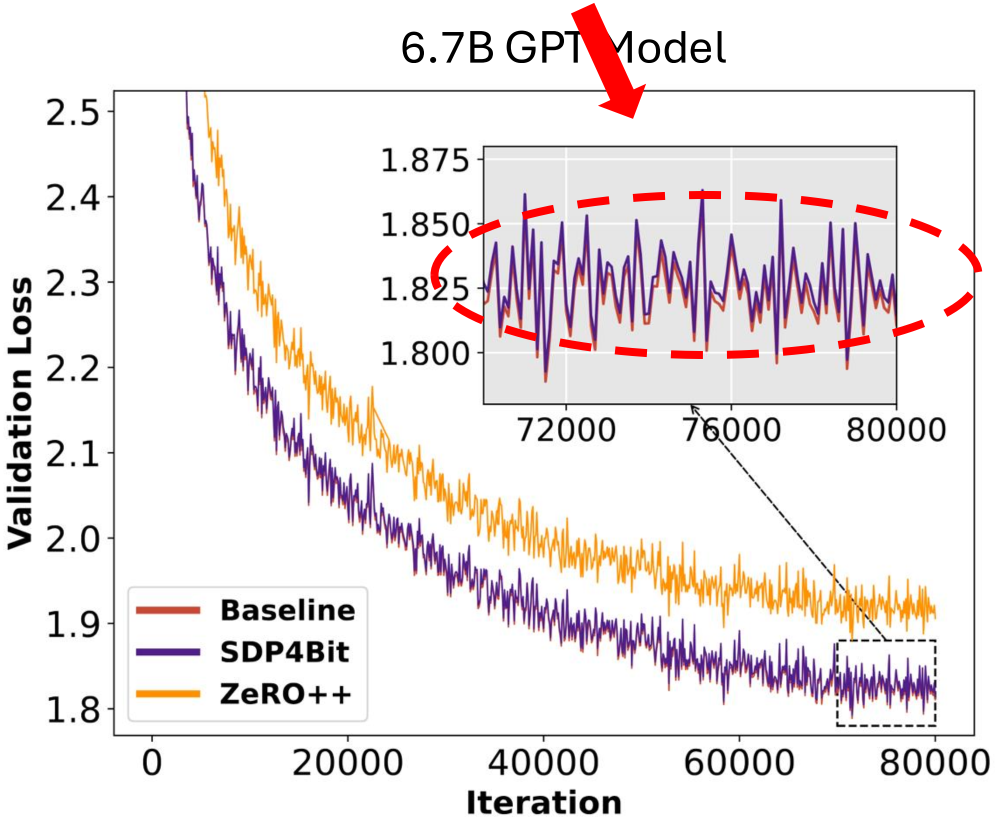
Intra Node: 32 bits → 8 bits *(can be hidden)*

Inter Node: 32 bits → 4 bits

4x reduction for weight

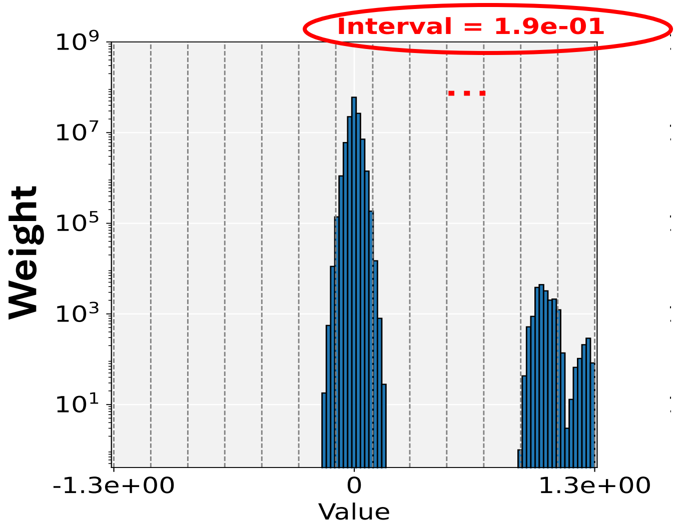
8x reduction for gradient

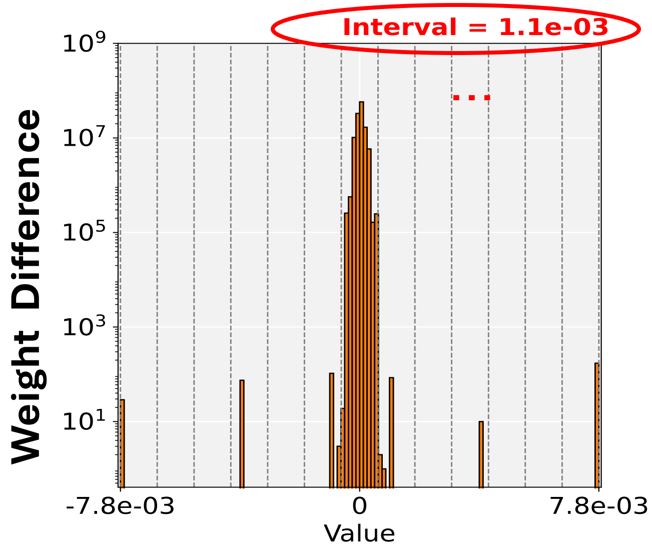
Almost no accuracy loss



INDIANA UNIVERSITY

Weight Compression Strategy



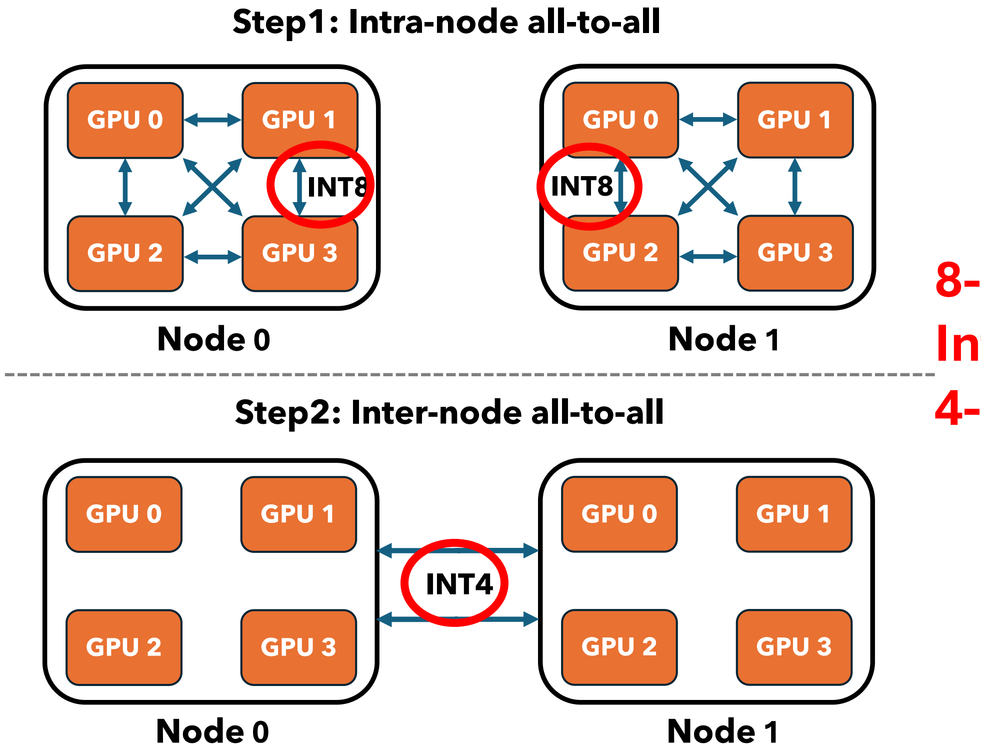


Weight Differences usually have a smaller range, so that can achieve a much lower compression error.

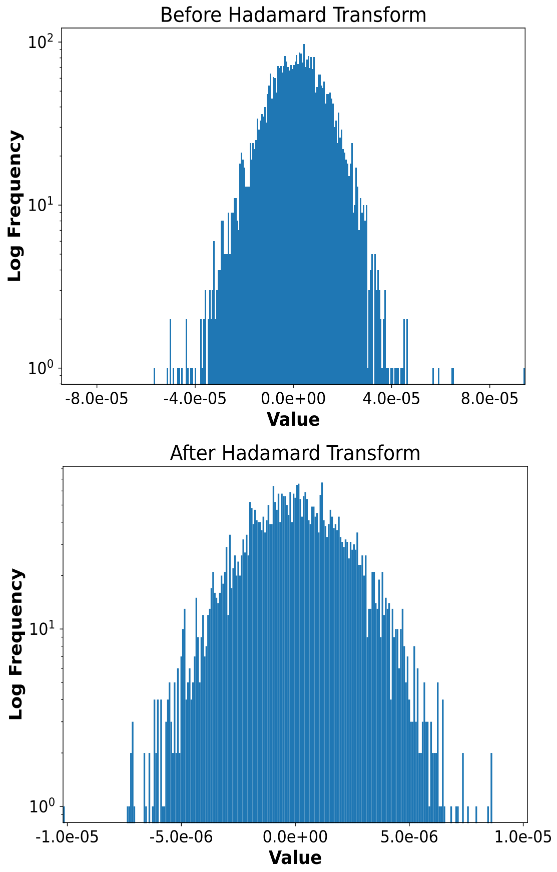
Use weight difference communication pattern, and quantize weight differences into 4 bits.

Gradient Compression Strategy

NEURAL INFORMATION PROCESSING SYSTEMS



8-bit quantization for Intra-node
4-bit for Inter-node

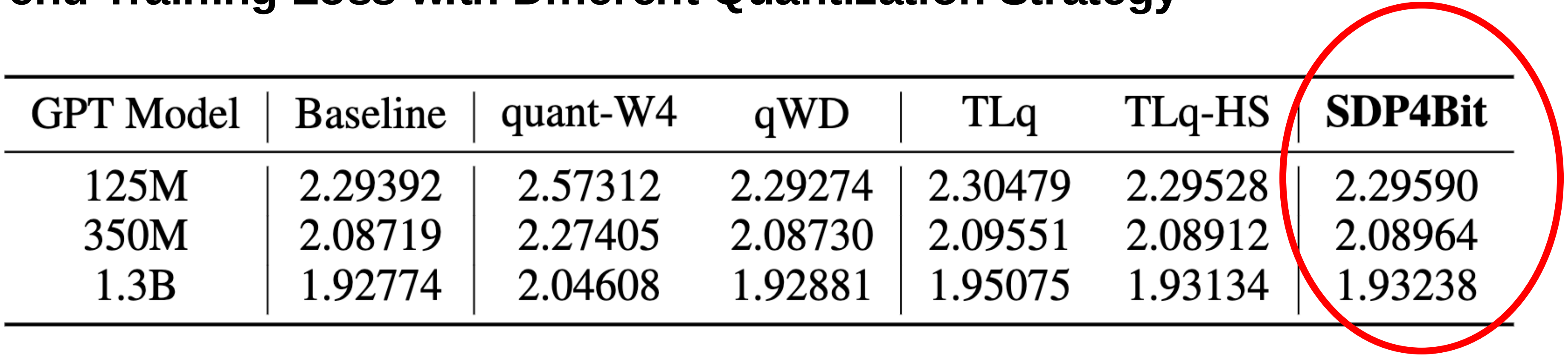


Using Hadamard Transformation to alleviate gradient outliers. (zero-overhead integration)

INDIANA UNIVERSITY

Experiments

- ➢ **End-to-end Training Loss with Different Quantization Strategy**



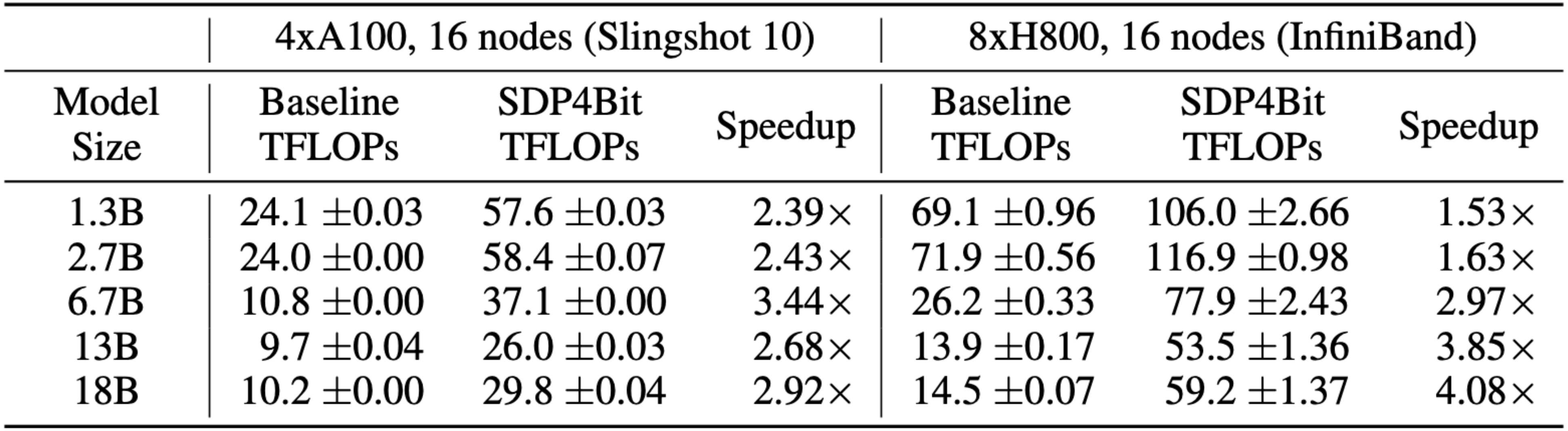
GPT Model	Baseline	quant-W4	qWD	TLq	TLq-HS	**SDP4Bit**
125M	2.29392	2.57312	2.29274	2.30479	2.29528	2.29590
350M	2.08719	2.27405	2.08730	2.09551	2.08912	2.08964
1.3B	1.92774	2.04608	1.92881	1.95075	1.93134	1.93238

Final loss is close to the baseline.

Experiments

NEURAL INFORMATION PROCESSING SYSTEMS

➢ E2E Training Throughput with Different Models



	4xA100, 16 nodes (Slingshot 10)			8xH800, 16 nodes (InfiniBand)		
Model Size	Baseline TFLOPs	SDP4Bit TFLOPs	Speedup	Baseline TFLOPs	SDP4Bit TFLOPs	Speedup
1.3B	24.1 ±0.03	57.6 ±0.03	2.39×	69.1 ±0.96	106.0 ±2.66	1.53×
2.7B	24.0 ±0.00	58.4 ±0.07	2.43×	71.9 ±0.56	116.9 ±0.98	1.63×
6.7B	10.8 ±0.00	37.1 ±0.00	3.44×	26.2 ±0.33	77.9 ±2.43	2.97×
13B	9.7 ±0.04	26.0 ±0.03	2.68×	13.9 ±0.17	53.5 ±1.36	3.85×
18B	10.2 ±0.00	29.8 ±0.04	2.92×	14.5 ±0.07	59.2 ±1.37	4.08×

Up to 4.08× Throughput Improvement.

INDIANA UNIVERSITY

NEURAL INFORMATION
PROCESSING SYSTEMS

Looking forward to see you on Dec 11

SDP4Bit: Toward 4-bit Communication Quantization in Sharded Data Parallelism for LLM Training

Poster Session 2

Wed 11 Dec 4:30 p.m. PST — 7:30 p.m. PST

INDIANA UNIVERSITY