NEURAL INFORMATION PROCESSING SYSTEMS

DUO: No Compromise to Accuracy Degradation

Jinda Jia , Cong Xie*, Fanjiang Ye, Hao Feng, Hanlin Lu, Daoce Wang, Haibin Lin, Zhi Zhang, Xin Liu*

INDIANA UNIVERSITY

ByteDance

* Equaly contributed first authors

Motivation:
Gradient Quantization Results in Accuracy Degradation

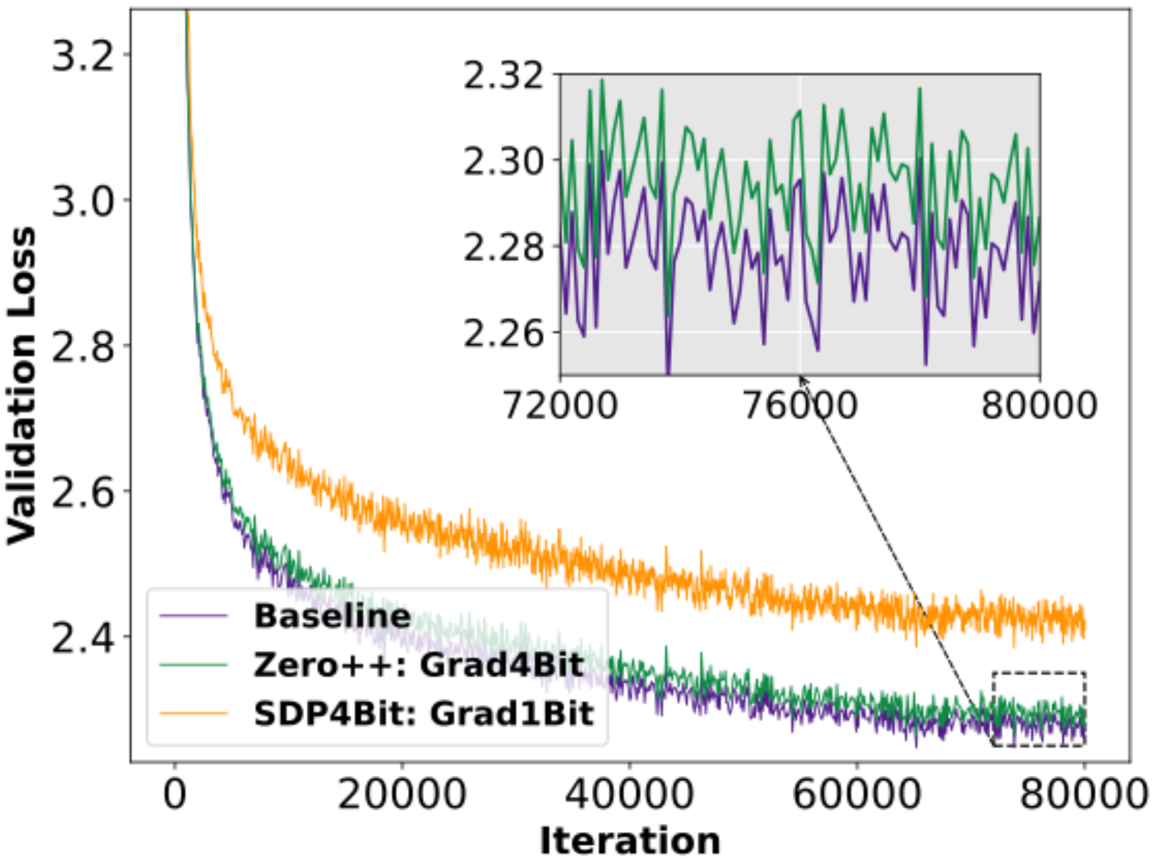
ZeRO++ with 4bit Gradient Quantization:

Intra/Inter Grad: 32 bits → 4 bits

SDP4Bit with 1bit Gradient Quantization:

Intra Node Grad: 32 bits → 8 bits

Inter Node Grad: 32 bits → 1 bits



Loss of ZeRO++, SDP4Bit

NEURAL INFORMATION PROCESSING SYSTEMS

DUO = Fast-Slow (algorithm) + Comp-Comm Overlapping + Gradient Offload

INDIANA UNIVERSITY

Design 1 Algorithm: Fast-Slow Reduction

NEURAL INFORMATION PROCESSING SYSTEMS

"Slow" Gradient Reduction:

1) High Precision FP32
2) Async with forward-backward
3) On the CPU

"Fast" Gradient Reduction:

1) Low Precision INT1/INT4/INT8...
2) Sync before Optimizer Step
3) On the GPU

Algorithm 1 Distributed training with DUO

Require: parameter weight (main copy): w, weight for forward-backward: $\tilde{w}$, weight for communication: w', weight difference: d, gradient: g, worker: p, pth shard of weight or gradient: $[p]$, gradient produced by worker p: g^p, gradients with compression $\tilde{g}$, averaged gradient: $\bar{g}$, current iteration: t

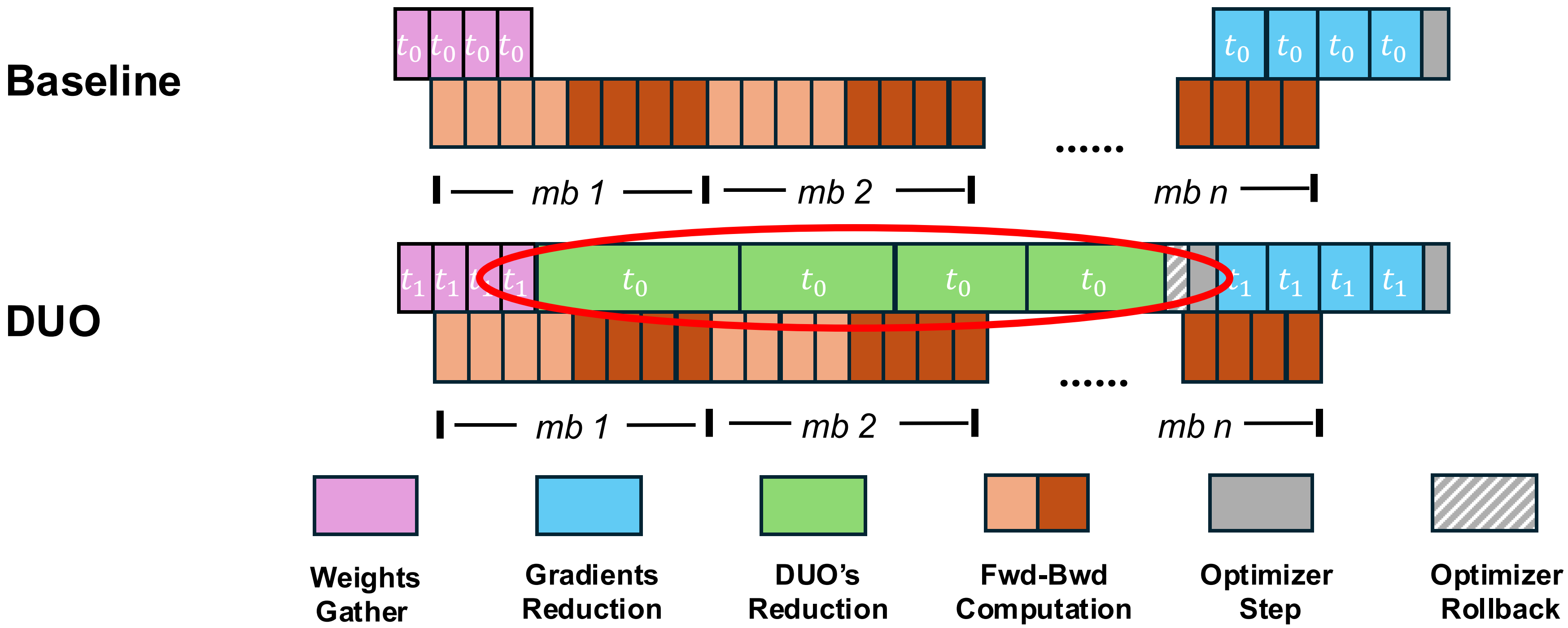
function ForwardPass
  $d_t[p] = w'_t[p] - \tilde{w}_{t-1}[p]$
  $\tilde{d}_t[p] \leftarrow \text{QuantizeWeightsDiff}(d_t[p])$
  $\tilde{d}_t \leftarrow \text{AllGather}(\tilde{d}_t[p])$
  Start $\text{AsyncReduceScatter}(g^p_{t-1})$
  $\tilde{w}_t \leftarrow \tilde{w}_{t-1} + \tilde{d}_t$
  $output^p \leftarrow \text{ForwardPass}(\tilde{w}_t, input^p)$
function BackwardPass
  $g^p_t \leftarrow \text{Gradient}(\tilde{w}_t, output^p_t)$
  $\text{SaveReplica}(g^p_t)$
  $\tilde{g}^p_t \leftarrow \text{CompressGradient}(g^p_t)$
  $\tilde{\bar{g}}_t[p] \leftarrow \text{ReduceScatter}(\tilde{g}^p_t)$
  $\bar{g}_{t-1}[p] \leftarrow$ Wait AsyncReduceScatter Finish
  Recover $w_{t-1}[p]$
  $w_t[p] \leftarrow \text{Optimizer}(\bar{g}_{t-1}[p], w_{t-1}[p])$
  $w'_{t+1}[p] \leftarrow \text{Optimizer}(\tilde{\bar{g}}_t[p], w_t[p])$

INDIANA UNIVERSITY

Design 2 System: Overlapping Communication with Computation

Existing *LLM training overlaps only with the first and last micro-batch.*

Ours *overlaps "slow" reduction with other idle phases.*



mb: micro-batch

NEURAL INFORMATION PROCESSING SYSTEMS

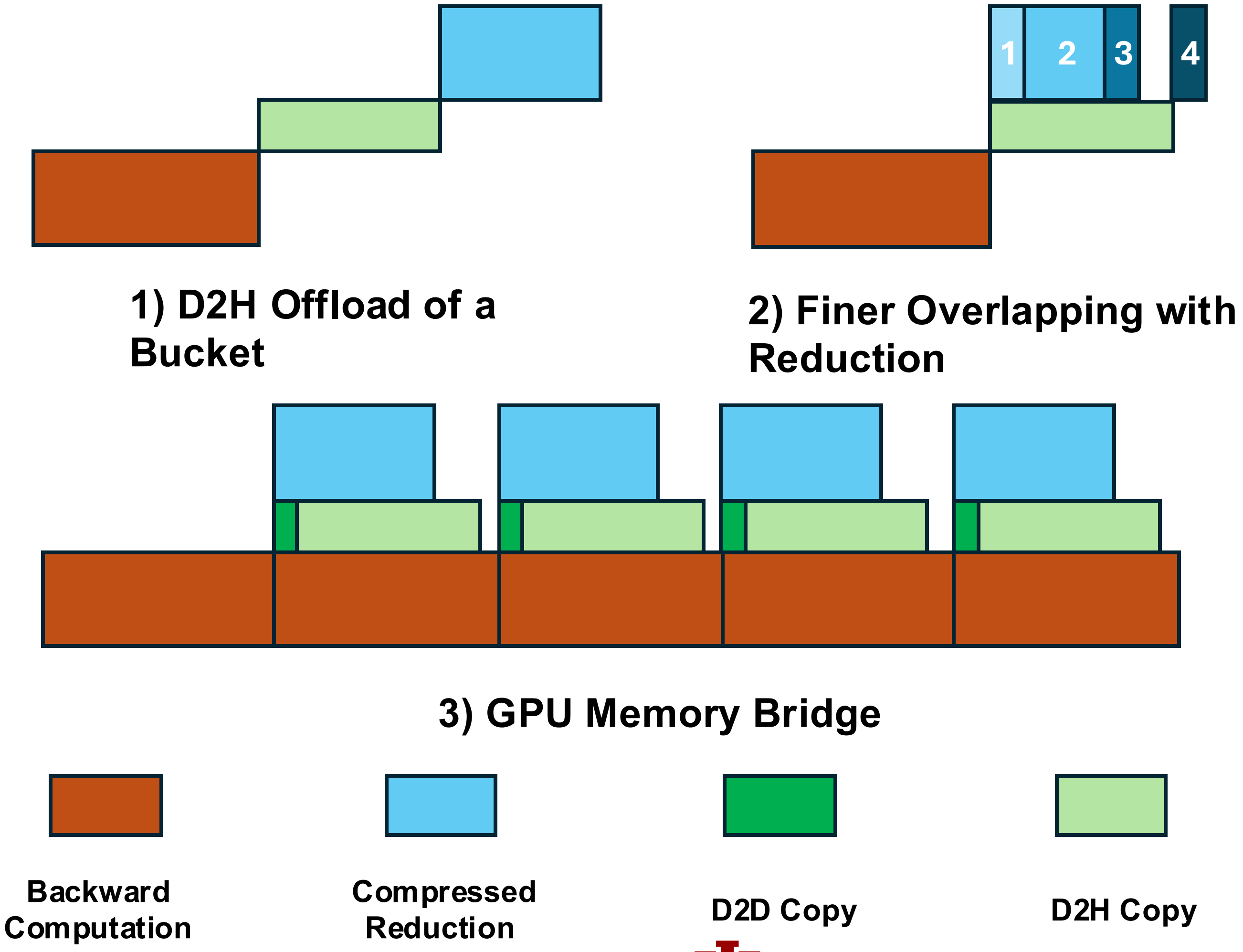
INDIANA UNIVERSITY

Design 3 System: Gradient Offload Optimization

NEURAL INFORMATION PROCESSING SYSTEMS

DUO requires copying gradients to CPU before starting the "fast" reduction (D2H copy bottleneck).

1) Bucket-wise offloading
2) Fine-grained D2H overlap
3) GPU memory bridge



INDIANA UNIVERSITY

Experiments

NEURAL INFORMATION PROCESSING SYSTEMS

➢ **End-to-end Training Loss with DUO**

Grad Bits	Strategy	125M	350M	1.3B	6.7B
32	Full Precision	2.2716	2.0582	1.8854	1.7527
4	SDP4Bit	2.2757	2.0629	1.8944	1.7570
	DUO	2.2727	2.0592	1.8907	1.7535
1	SDP4Bit	2.4204	2.1843	2.0489	1.8434
	DUO	2.2712	2.0686	1.8943	1.7572
0	**DUO**	2.2761	2.0628	1.8941	1.7550

Final loss is close to the baseline. Even fully remove "fast" reduction (0-bit)

INDIANA UNIVERSITY

Experiments

NEURAL INFORMATION PROCESSING SYSTEMS

➢ E2E Training Throughput with Different Models

Model	Method	H20 (8×8 H20 Nodes)		A100 (8×4 A100 Nodes)	
		Throughput (TFLOPS)	Mem (MB)	Throughput (TFLOPS)	Mem (MB)
1.3B	DUO	70.82 ± 0.63	15880	117.53 ± 0.49	21270
	SDP4Bit	71.02 ± 0.63	15832	117.61 ± 2.35	20934
2.7B	DUO	78.71 ± 0.92	25600	125.31 ± 0.37	35298
	SDP4Bit	79.15 ± 0.80	25640	126.42 ± 0.82	34422
6.7B	DUO	84.09 ± 2.04	16162	122.67 ± 3.76	26480
	SDP4Bit	85.61 ± 1.57	15692	121.05 ± 1.89	25684
13B	DUO	93.02 ± 1.08	27596	132.88 ± 0.95	27332
	SDP4Bit	94.68 ± 0.77	27220	136.73 ± 1.35	27142
18B	DUO	100.85 ± 1.29	36540	116.37 ± 1.28	27378
	SDP4Bit	102.93 ± 1.04	35478	122.91 ± 1.16	27088

DUO achieves similar throughput to SDP4Bit, showing that our high-precision communication adds minimal overhead.

INDIANA UNIVERSITY

Looking forward to see you on Dec 3

NEURAL INFORMATION PROCESSING SYSTEMS

DUO: No Compromise to Accuracy Degradation

Exhibit Hall C,D,E
Wed 3 Dec 4:30 p.m. PST — 7:30 p.m. PST

INDIANA UNIVERSITY