DRAGONN: Distributed Randomized Approximate Gradients of Neural Networks

Zhuang Wang*, Zhaozhuo Xu*, Xinyu Crystal Wu, Anshumali Shrivastava and T. S. Eugene Ng

* equal contribution

Communication Bottleneck in DDT

Computation gets faster

- Advanced DNN accelerators
    - P100 -> V100 -> A100
- Advanced DNN compilers
    - XLA, TVM, etc.
- The single-GPU iteration time of ResNet50 has improved by ~22x

Network bandwidth can’t catch up

- Slower-growing bandwidth
    - ~10x increase

Communication becomes the performance bottleneck

System View of Gradient Sparsification (GS)

Top-k gradients for synchronization

- Exact TopK GS
- Approximate TopK GS, e.g., DGC

Save up to 99.9% gradient exchange

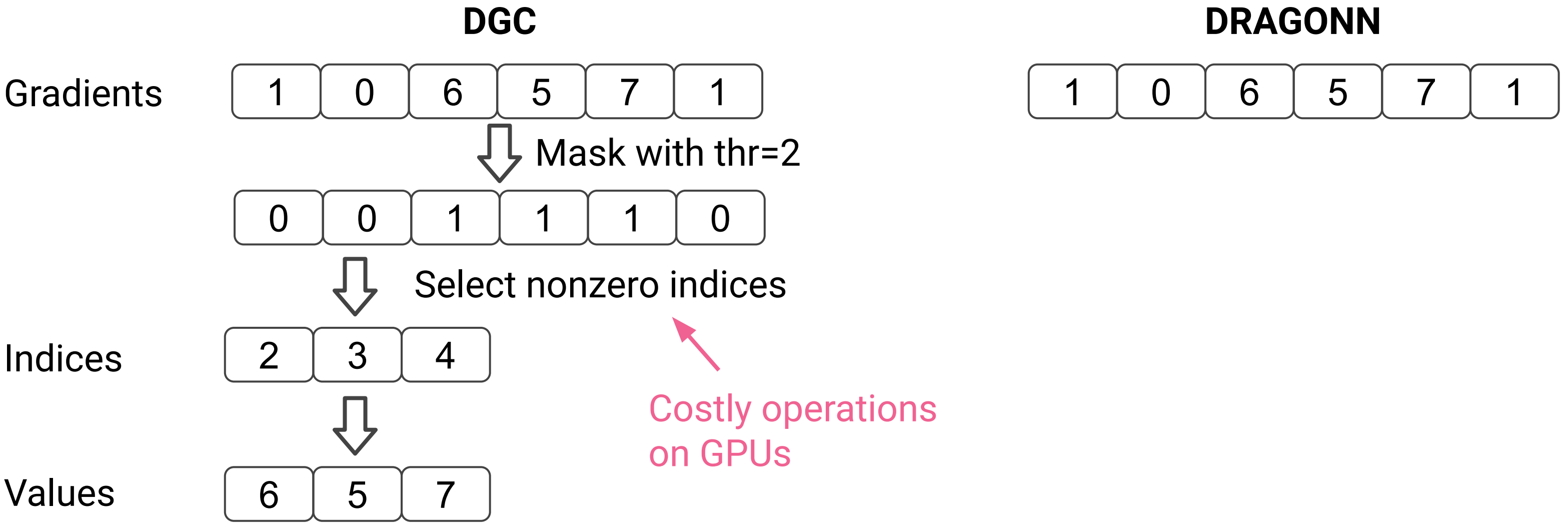
- Greatly reduce communication time

Previous work looked at GS from a theoretical perspective

- They ignore the high cost of sparsification
- GS computation time can exceed communication time
- Lead to limited end-to-end improvements

DRAGONN: encoding

Cheap encoding operations



DGC

DRAGONN

Gradients

1	0	6	5	7	1

Mask with thr=2

0	0	1	1	1	0

Select nonzero indices

Indices

2	3	4

Values

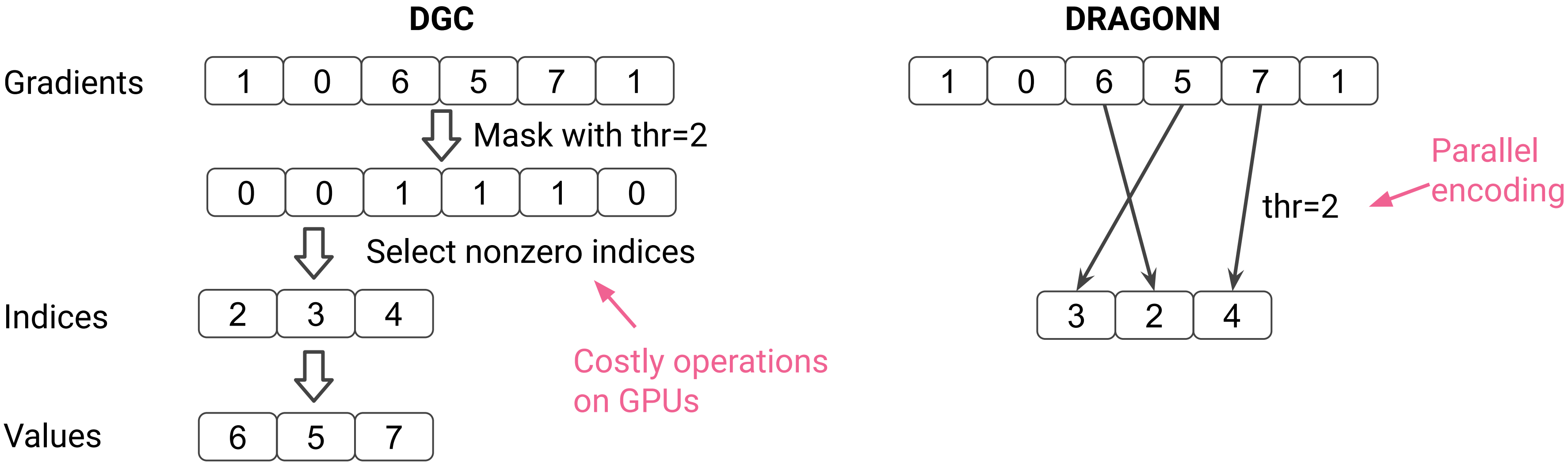
6	5	7

1	0	6	5	7	1

Costly operations on GPUs

DRAGONN: encoding

Cheap encoding operations



DGC

Gradients

1	0	6	5	7	1

Mask with thr=2

0	0	1	1	1	0

Select nonzero indices

Indices

2	3	4

Values

6	5	7

Costly operations on GPUs

DRAGONN

1	0	6	5	7	1

thr=2

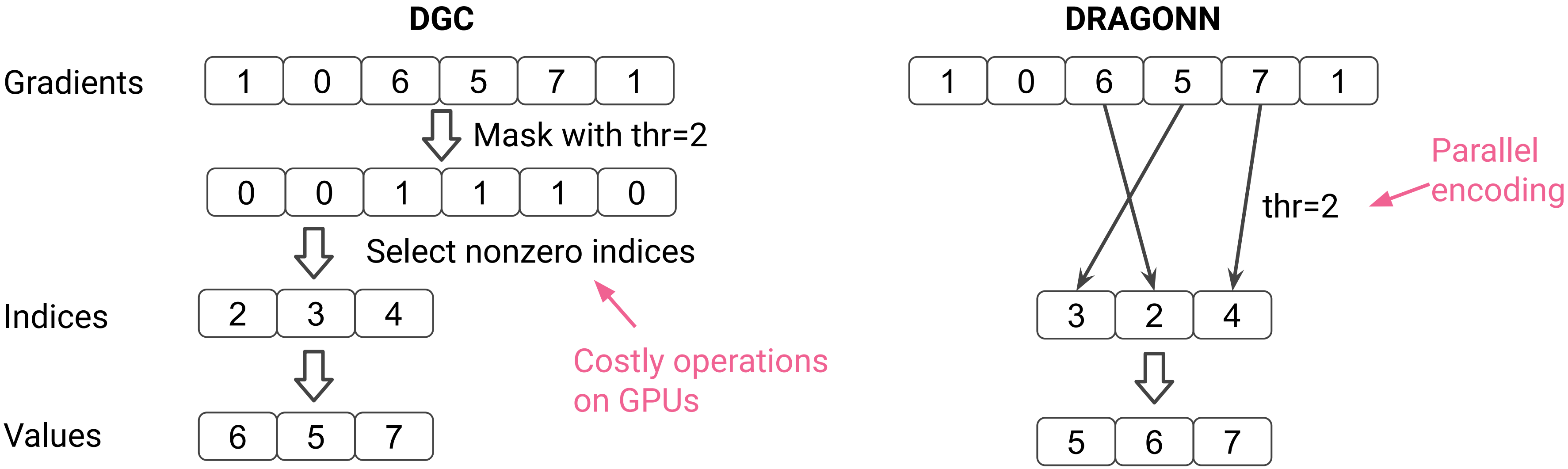
3	2	4

Parallel encoding

DRAGONN naturally supports massively parallel encoding

DRAGONN: encoding

Cheap encoding operations



DGC

Gradients: 1 0 6 5 7 1

Mask with thr=2

0 0 1 1 1 0

Select nonzero indices

Indices: 2 3 4

Values: 6 5 7

Costly operations on GPUs

DRAGONN

1 0 6 5 7 1

thr=2

Parallel encoding

3 2 4

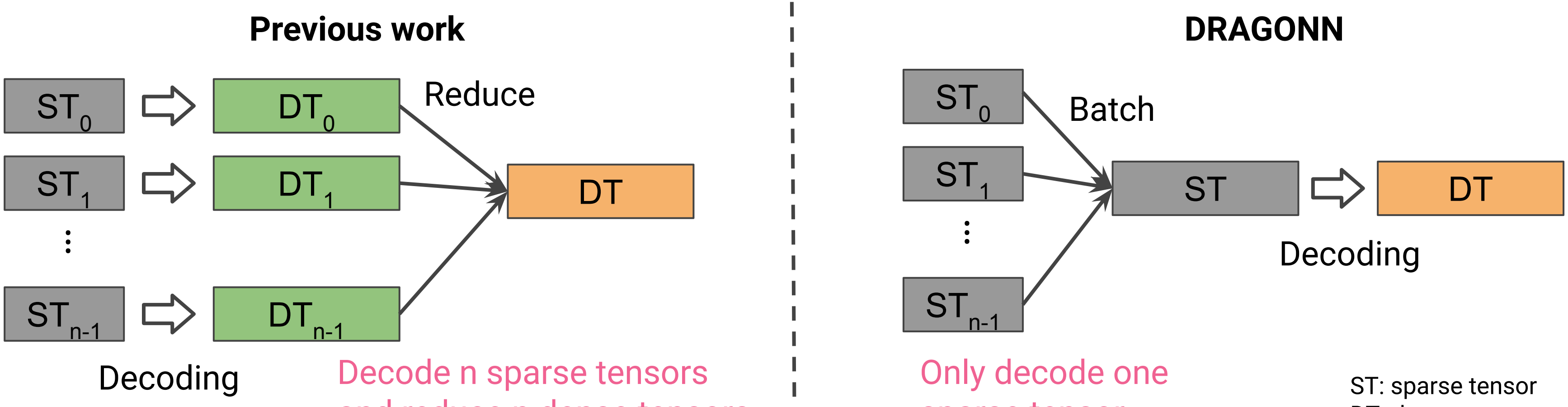
5 6 7

DRAGONN naturally supports massively parallel encoding

DRAGONN: decoding

Cheap decoding operations

- Index-value pairs are independent of each other
- Near-constant decoding time regardless of the number of GPUs



Decode n sparse tensors and reduce n dense tensors

Only decode one sparse tensor

ST: sparse tensor
DT: dense tensor

DRAGONN: tensor selection

Efficiency-aware tensor selection for GS

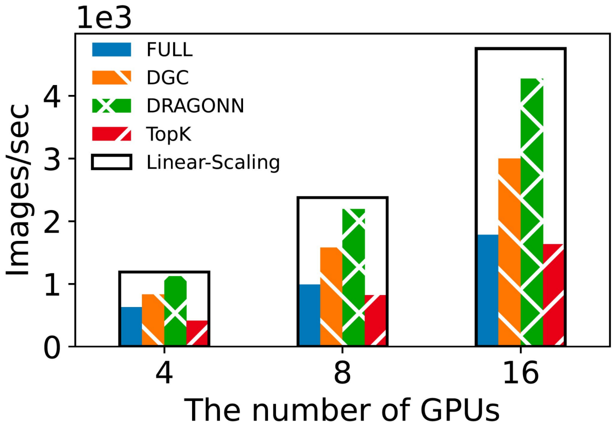
- A general cost-benefit analysis based on offline profiling

$$T_{\text{comp}}(d) < T_{\text{full}}(d) - T_{\text{spr}}(d)$$

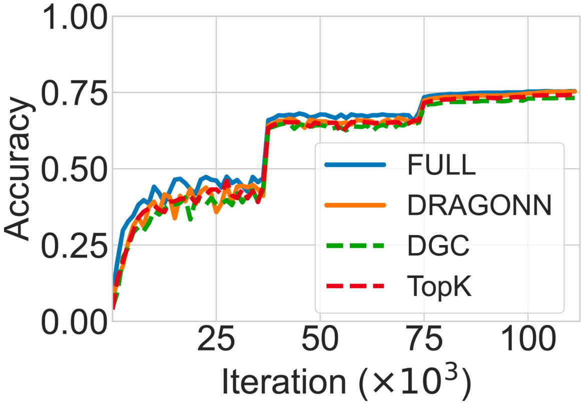
- Apply DRAGONN to tensors only when it benefits the iteration time

Evaluations

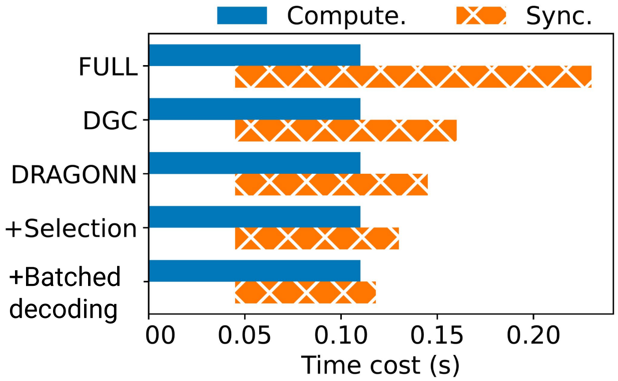
ResNet50 over ImageNet



Training throughput



Accuracy vs. Iteration



Improvement breakdown

Summary

- **Measurements** to understand the real world GS overheads
- DRAGONN is the first work to address this challenge with a **randomized hashing algorithm**
- **Theoretical analysis** on DRAGONN
- It significantly **reduces the encoding and decoding overheads**, while preserving the iteration wise accuracy

Contact: zhuang.wang@rice.edu