ETL: Elastic Training Layer for Deep Learning

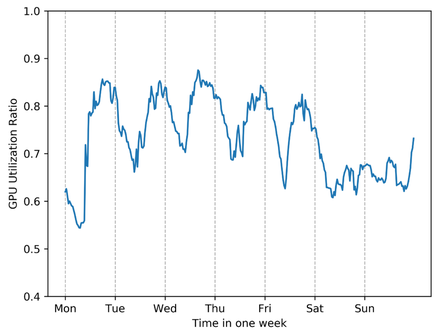
Lei Xie, Jidong Zhai
Tsinghua University

Introduction & Motivation

1. Elastic training benefits deep learning clusters

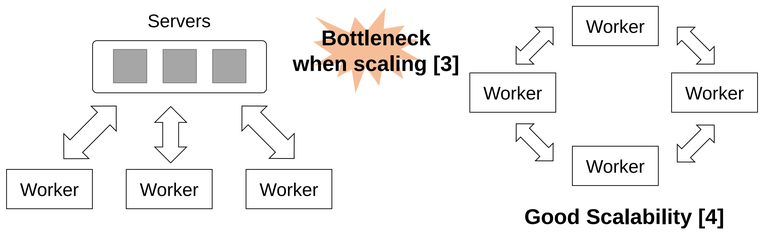
- Leverage available resources during spare time
- Migrate jobs to avoid interference and defragmentation

GPU Utilization of a DL cluster

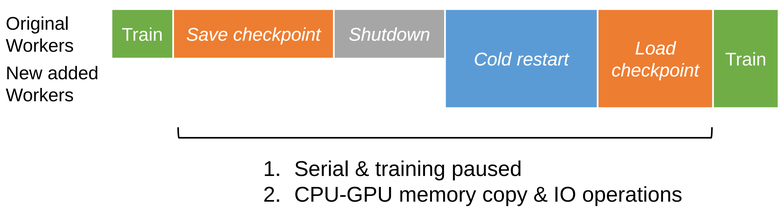


- Gandiva [1], Optimus [2]

2. Existing elastic training is based on low performance parameter-server architecture



3. Existing elastic training is based on checkpoint



1. Serial & training paused
2. CPU-GPU memory copy & IO operations

References:
[1] Xiao, Wencong, et al. "Gandiva: Introspective cluster scheduling for deep learning." *13th USENIX Symposium on Operating Systems Design and Implementation (OSDI 18)*. 2018.
[2] Peng, Yanghua, et al. "Optimus: an efficient dynamic resource scheduler for deep learning clusters." *Proceedings of the Thirteenth EuroSys Conference*. ACM, 2018.
[3] Ma, Minghuang, et al. "Democratizing Production-Scale Distributed Deep Learning." *arXiv preprint arXiv:1811.00143* (2018).
[4] Kurth, Thorsten, et al. "Exascale deep learning for climate analytics." *Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis*. IEEE Press, 2018.

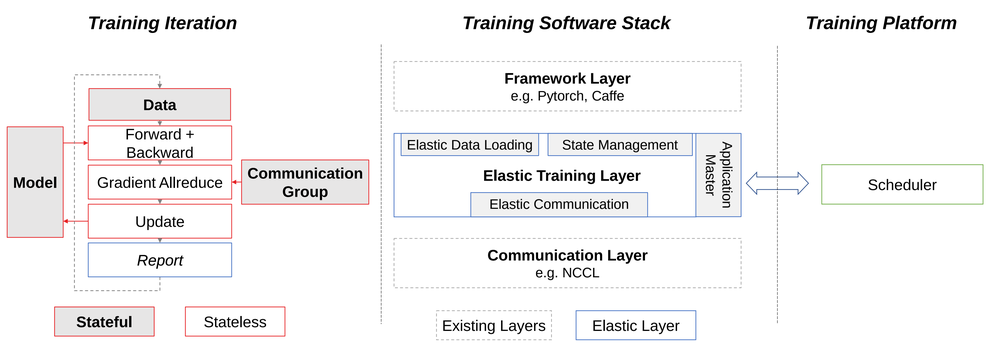
ETL: Elastic Training Layer

Challenges

- High performance communication with low overhead elasticity
- Heterogeneity from accelerators, e.g. GPU
- Efficient state replication

Targeting for

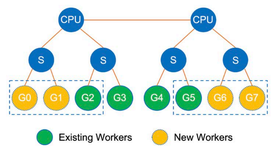
- Synchronous distributed training
- Data parallelism
- Gradient aggression based on Allreduce primitive



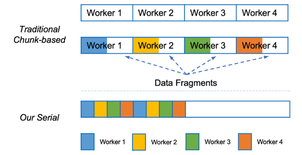
Mechanisms

- Configurable *report* stage
- Serial data loading pipeline
- Asynchronous, parallel and IO-free state replication

State replication mechanism



Serial data loading pipeline



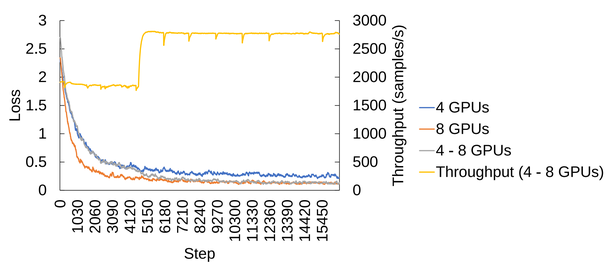
Evaluation

Environment

- Hardware: Machines with 8 NVIDIA 1080Ti GPUs on each
- Software: Elastic training with Pytorch and NCCL
- Task: ResNet50 training on Cifar10

Showcase: Elastic Training of ResNet50 on Cifar10

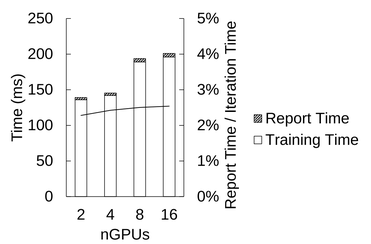
- Static training on 4 GPUs
- Static training on 8 GPUs
- Elastic training on 4 GPUs and then 8 GPUs after 6 epochs



More GPU resources yield higher training throughput and thus leads to higher training efficiency.

Overhead of Elasticity

- Overhead from the lightweight and configurable report stage
- An iteration consists of training and report



Overhead of *report* is less than 2.5% of iteration time, and is quite stable across 2-16 GPUs.

Efficiency of Replication

- 4 GPUs to 8 GPUs

Phase		Time (s)
Original 4	New 4	
Save checkpoint		0.245
Shutdown		5.465
Cold restart		64.883
Load checkpoint		2.722
Train		
Total		73.315

Phase		Time (s)
Original 4	New 4	
Train	Cold start	62.374 (hidden)
Report & replication		1.065
Train		
Total		1.065

Email: xlei.thu@gmail.com