Analysis of Large-Scale Multi-Tenant GPU Clusters for DNN Training Workloads

Myeongjae Jeon, Shivaram Venkataraman, Amar Phanishayee,

Junjie Qian, Wencong Xiao, Fan Yang

Microsoft

UNIST
ULSAN NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

北京航空航天大學
BEIHANG UNIVERSITY

Deep Learning at a Large Enterprise

Speech, Image, Ads, NLP, Web Search …

DL training jobs require large GPU clusters

Philly: *Cluster manager for DL workloads on large shared GPU clusters*

		Motivated by observations in Philly	
Recent Cluster Managers	**Optimus [EuroSys 18]**	**Gandiva [OSDI 18]**	**Tiresias [NSDI 19]**
Objective	Average JCT	Consolidation	Average JCT
Scheduler	SRTF	Time-sharing	Gittins Index

Contributions

1. First characterization study of large-scale GPU clusters for DNN training

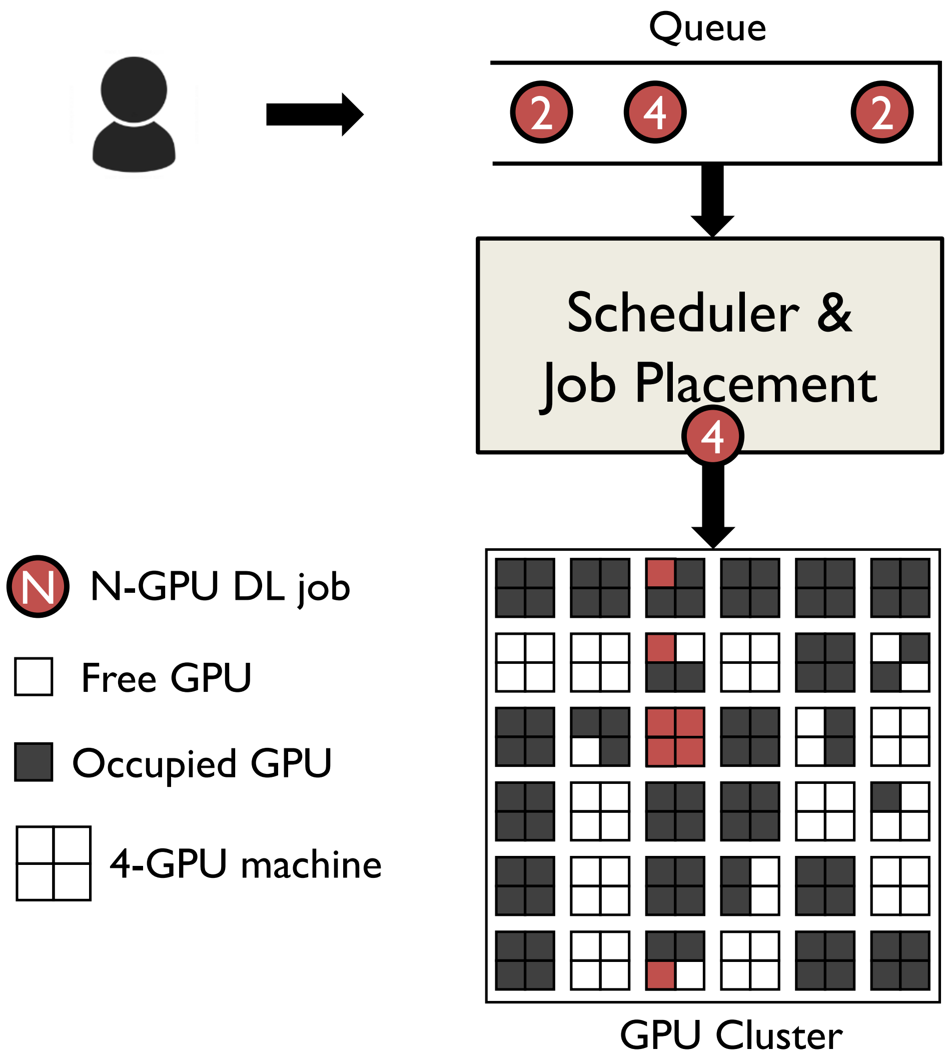
75-day period from Oct. 2017 to Dec. 2017

Total of 96,260 jobs over 14 virtual clusters

2. Uncover inefficiencies in cluster utilization

3. Present lessons for better cluster manager designs

Low Locality in Distributed Training



High intra-server locality

- High communication efficiency
- Long waiting time

Low intra-server locality

- Low waiting time
- Contention in the use of network
- Risk of intra-server interference (across jobs)

Low GPU locality is detrimental to cluster efficiency → Prioritize locality

Failures across Stack Reduce Cluster Utilization

Infrastructure

Resource Scheduler

HDFS

AI Engine

PyTorch

TensorFlow

mxnet

User

```
class Net(nn.Module):
    def __init__(self):
        super(Net, self).__init__()
        self.conv1 = nn.Conv2d(1, 10, kernel_size=5)
        self.conv2 = nn.Conv2d(10, 20, kernel_size=5)
        self.conv2_drop = nn.Dropout2d()
        self.fc1 = nn.Linear(320, 50)
        self.fc2 = nn.Linear(50, 10)
```

Failure types

- **Frequency User errors in code**
- **Temporal Infrastructure failures**
- **↑GPU hours wastes Semantics errors**

Improve failure handling (e.g., pre-run jobs)

Analysis of Large-Scale Multi-Tenant GPU Clusters for DNN Training Workloads

Microsoft

UNIST
ULSAN NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

北京航空航天大學
BEIHANG UNIVERSITY

11:05 AM, Session - Scheduling Things (Track II),

on July 12 at USENIX ATC 2019