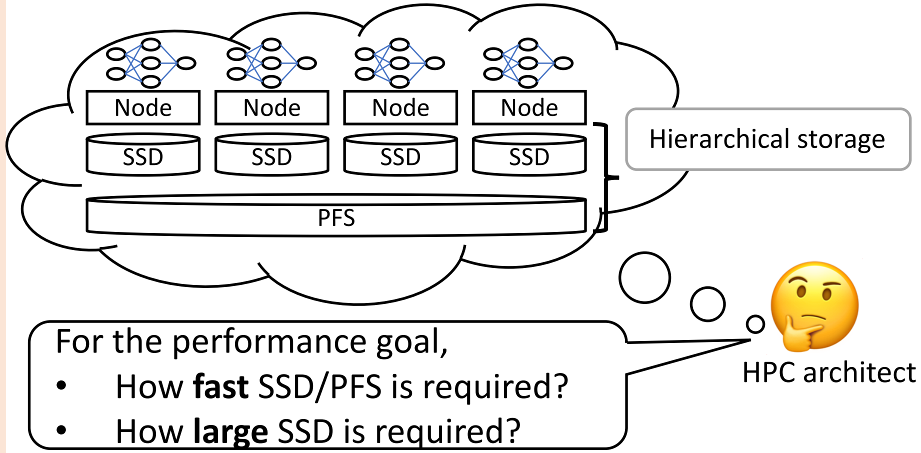
# Performance Measurement of a Hierarchical File System for Distributed Deep Neural Network Training

Takaaki Fukai, Kento Sato (RIKEN R-CCS)

## Background

- HPC is used for training deep neural network model with large network and dataset.
- In the workload, I/O performance is a critical matter.

In designing a HPC cluster for deep learning...



For the performance goal,

- How **fast** SSD/PFS is required?
- How **large** SSD is required?

## Research Goal

Goal: Analyzing the training performance with various storage system.

- How does the SSD/PFS throughput affect the training performance?
- How does the proportion of data on SSD to the dataset affect the training performance?

## Research Plan

We will measure training performance with various size and <u>speed</u> of the hierarchical storage.

Question: <u>Can we prepare the various “speed” of the storage in a cluster?</u>

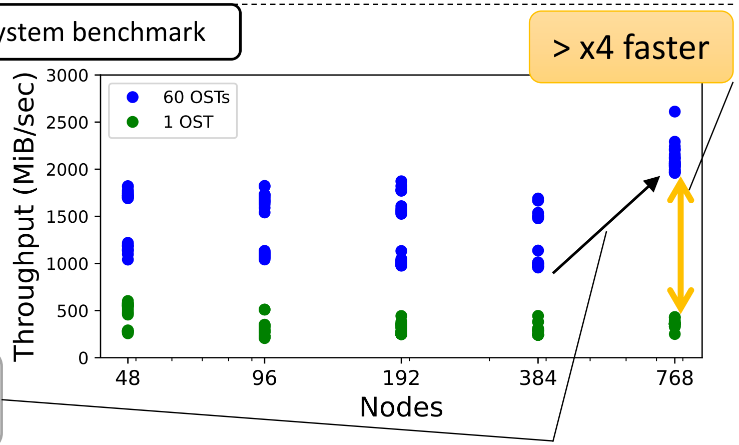
## Pre-experiments: Changing PFS performance on DL workload

Hypothesis: Different number of object storage targets (OST) of PFS offers different performance even on the DL workload.

- We execute two benchmarks to measure the performance on the global file system of supercomputer Fugaku with 1 OST and with 60 OSTs.
- The PFS is FEFS, a Lustre based PFS allowing us to control # of OST by lfs command.

### fio: a general file system benchmark

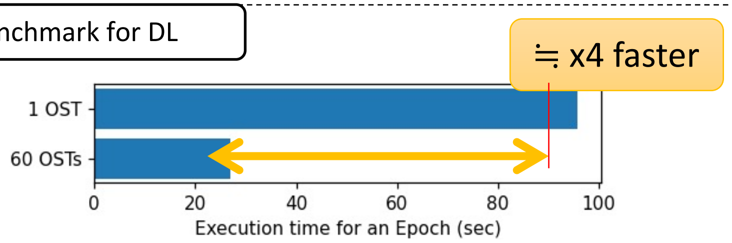
- We setup fio to read files in random order and read each file sequentially.
- We execute it on 48, 96, 192, 384, and 768 nodes.
- 8 processes run on each node.
- File size is 128KB, # of files/node is 256.



Throughput on 60 OSTs increases because the number of nodes connected to PFS is increased.

### DLIO: an I/O benchmark for DL

- DLIO simulates the I/O pattern in DL (e.g. shuffle, mini-batch)
- File type is tfrecord, local batch size is 12.
- File size is 128KB, and # of file is 576K (768 / nodes).



## Conclusion and Future works

Changing the number of OSTs of the PFS allow us to change the I/O throughput on DL training workload.

Our future works are

- Measuring the training performance with the 1OST/60OSTs setup and various DL workload
- Analyzing the measurement results and build performance model.