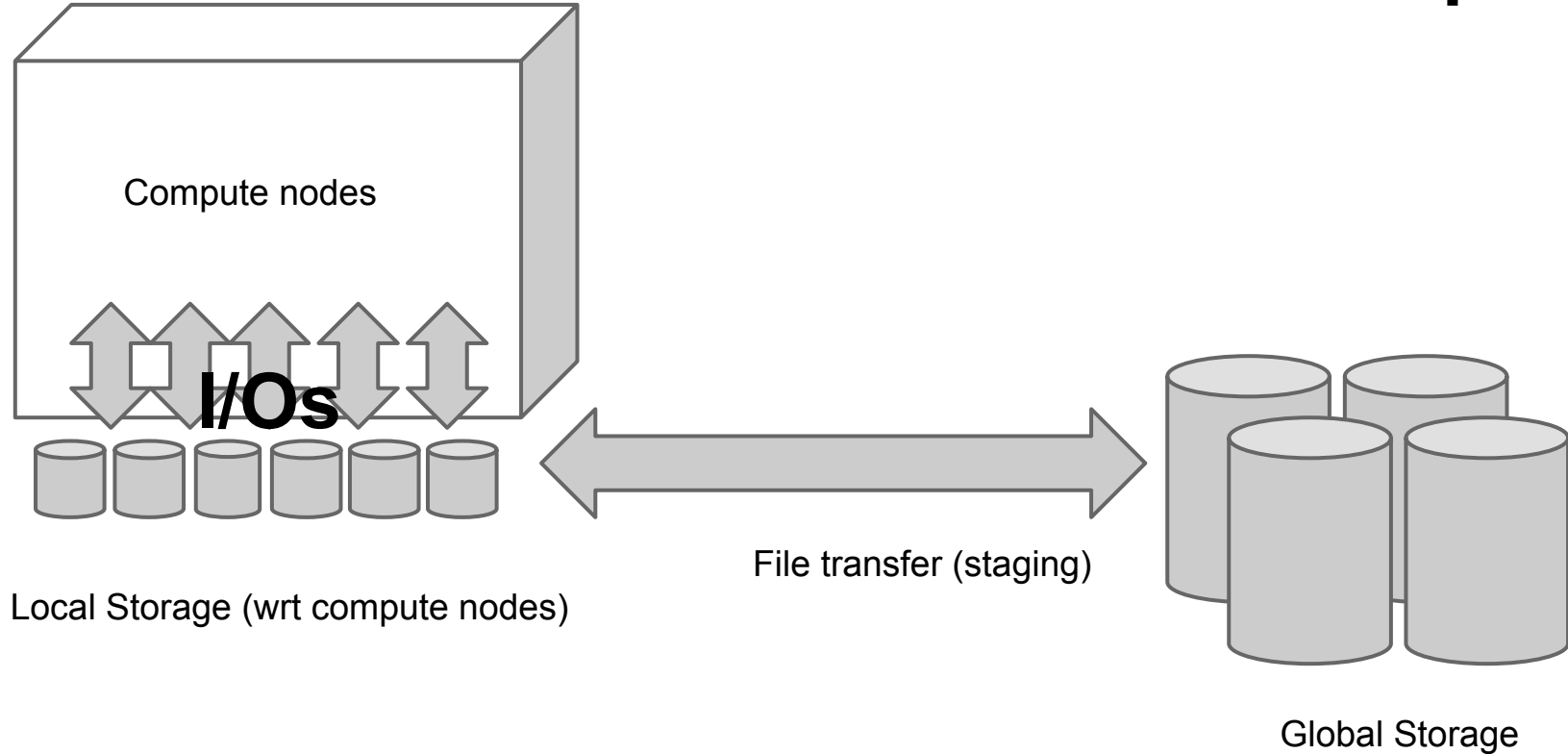# A User's Experience with FEFS

Itaru Kitayama
RIKEN AICS

Outline

- User's View of the FEFS for the K computer
- User's Notes on FEFS
- The Target MPI Application
- Effects of 3D-shape of Processes
- Towards 80,000 Nodes Parallelism
- Summary

User's View of the FEFS for the K computer

Compute nodes

I/Os

Local Storage (wrt compute nodes)

File transfer (staging)

Global Storage

User's Notes on FEFS

- File transferring to/from the compute nodes is mandatory (for the K computer)
  - Bring files to nodes, only 12 OSTs are provided, concerns for large datasets
  - At run-time IO applications see is up to 1 TB/s
- Rank directories (cut MDS access using loop-devices for MPI processes)
- Metadata accesses are good (<1,000 files in a dir)

The Target MPI Application

- DNA sequence analysis program ported to the K computer system
    - ABySS https://github.com/bcgsc/abyss
- IO characteristics of the App
    - Large analysis data read-in in the beginning
    - Result file write-outs (light IO) in the end

Running Time with Various Nodes (Shape)

Running Time min.
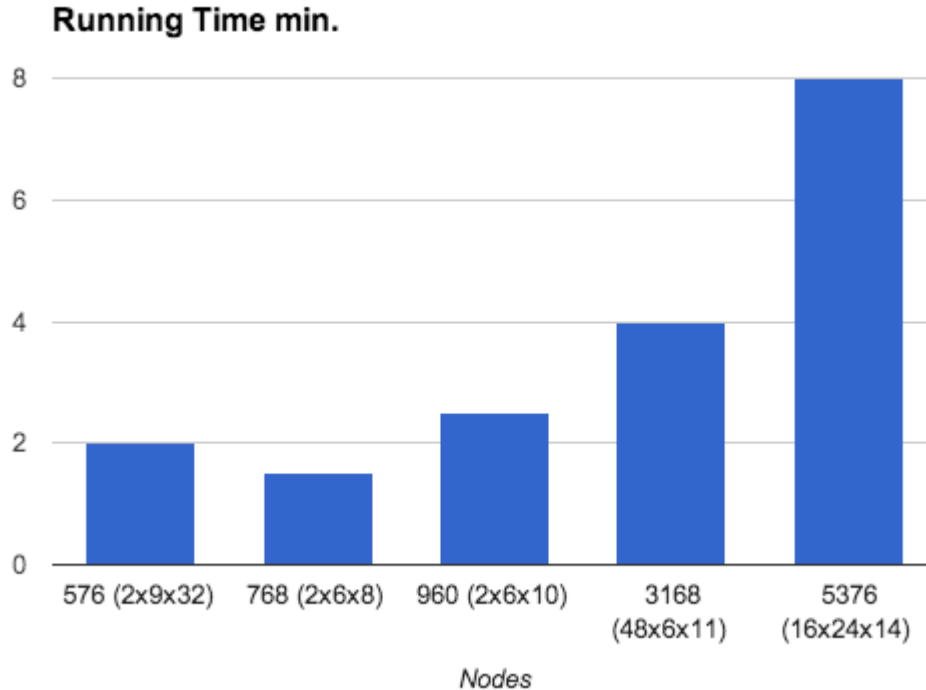
Need investigation what the best Job "3D-shape" for this App

Why Considering 3D Shape Matters

OSS's view of the nodes

Z=32

Compute node

OSS

Z=1

Z=0
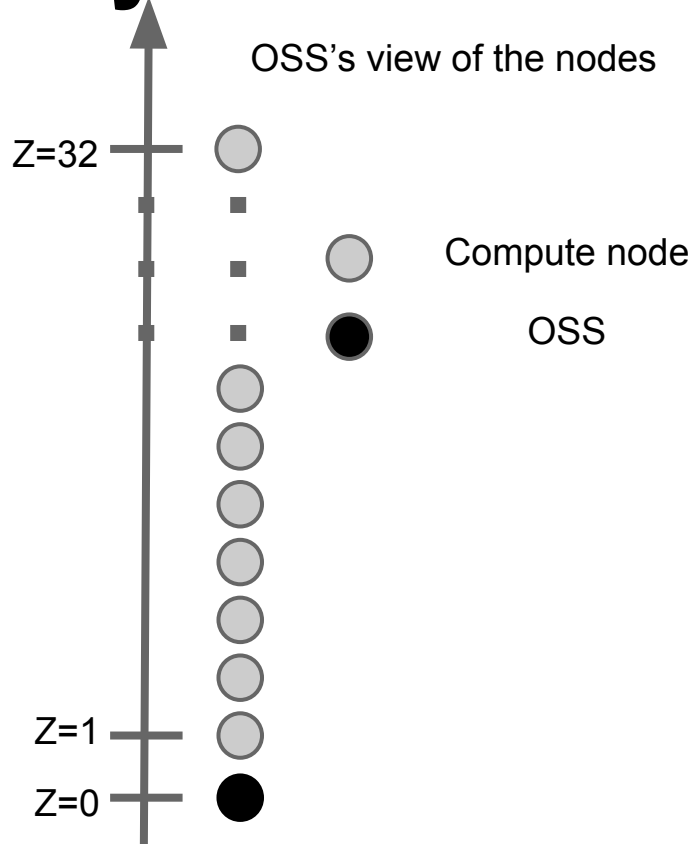
Excluding other user's jobs from the same Z-axis might help unexpected overloads of the OSS

Towards 80,000 Nodes Parallelism

- Always occupy the Z-axis to exclusively use OSSes (Trade off: scheduling gets tight)
- Sampling the running time with 196-node set-up (2x3x32 fixed job-shape)

Summary

- FEFS provides slightly different views on disks (local and global disks)
- Understanding job's shape is critical
- Higher performance (with 10,000 nodes or more) can be achieved by exclusive use of OSSes

Thanks to

Toshiyuki Maeda, RIKEN AICS
Shuichi Ihara, DDN
Shinji Sumimoto, Fujitsu

Exascale IO Workshop

Date: Dec 4, 2013

Location: RIKEN AICS, Kobe Japan

Topics: Storage systems for the Exascale era