Performance Comparison of DAOS and Lustre for Object Data Storage Approaches

Nicolau Manubens

Tiago Quintino

Simon Smart

ECMWF

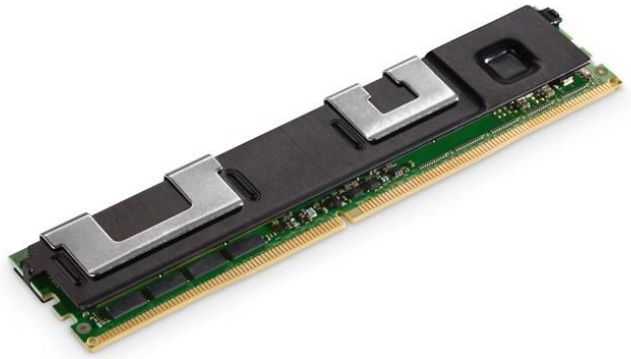
Adrian Jackson

a.jackson@epcc.ed.ac.uk

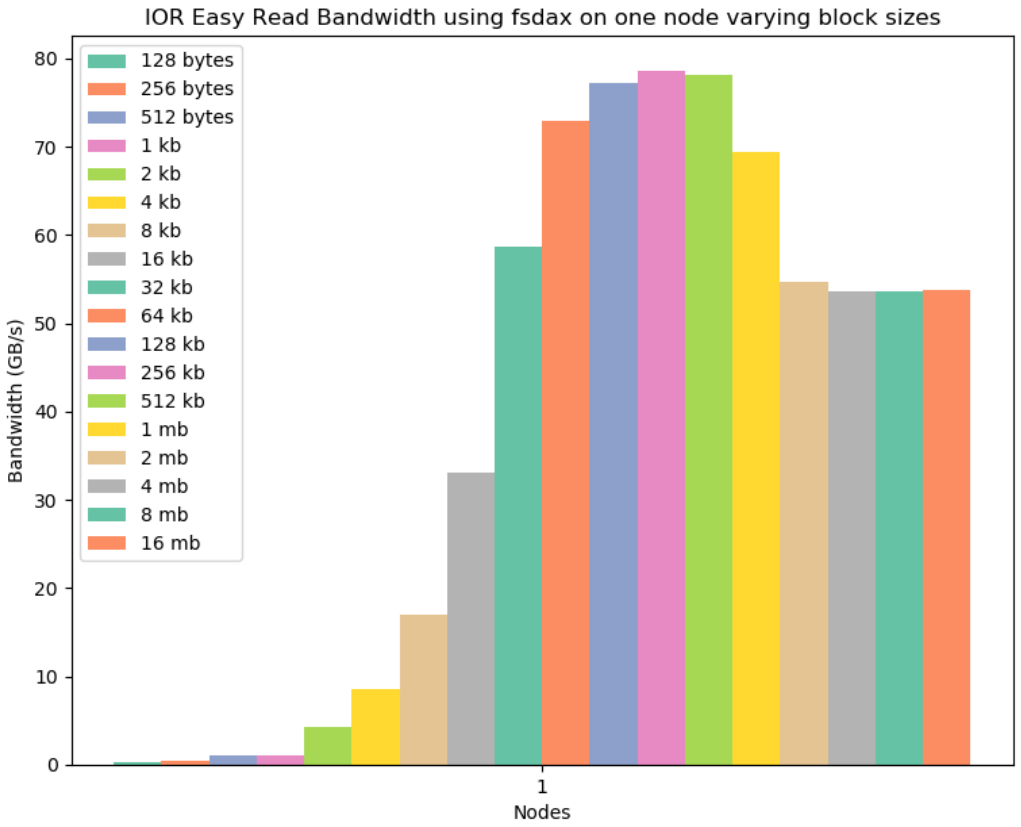
EPCC, The University of Edinburgh

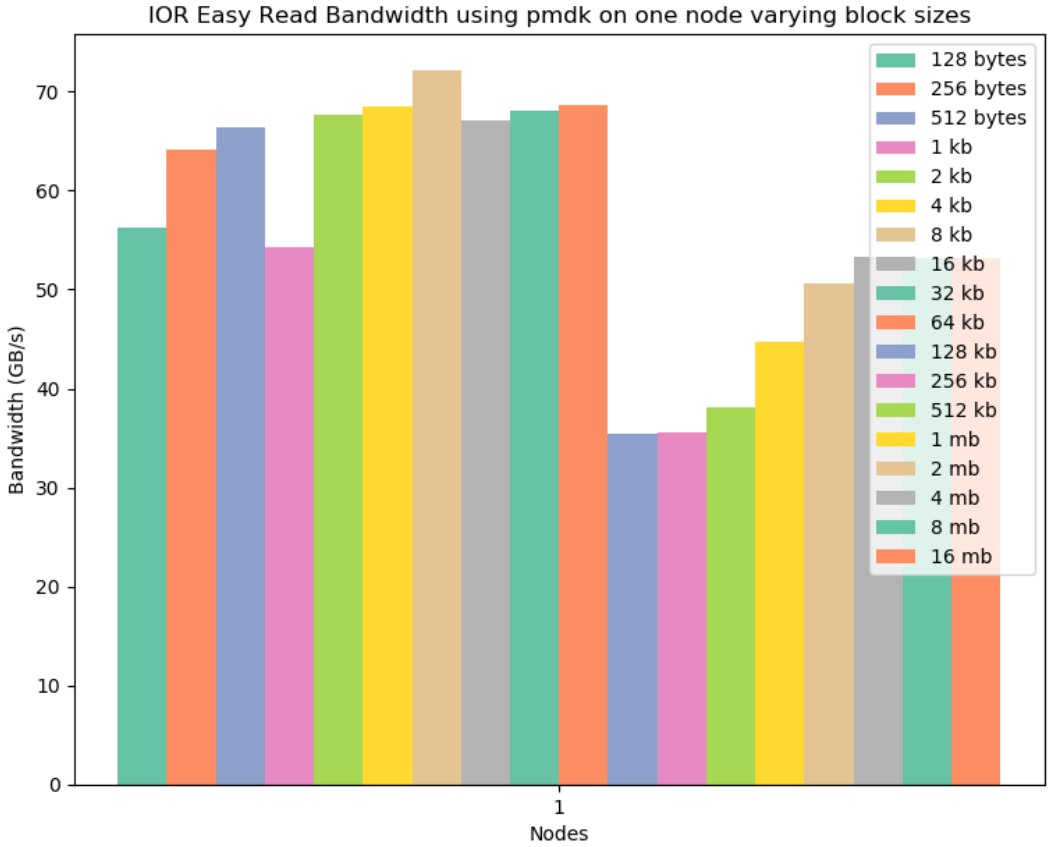
Storage hardware

- Non-volatile memory hardware
  - Optane
- Provides high random-access performance
  - High bulk I/O
- Localised I/O performance
- Potential replacements
  - NVMe
  - CXL disaggregated memory
  - Battery backup

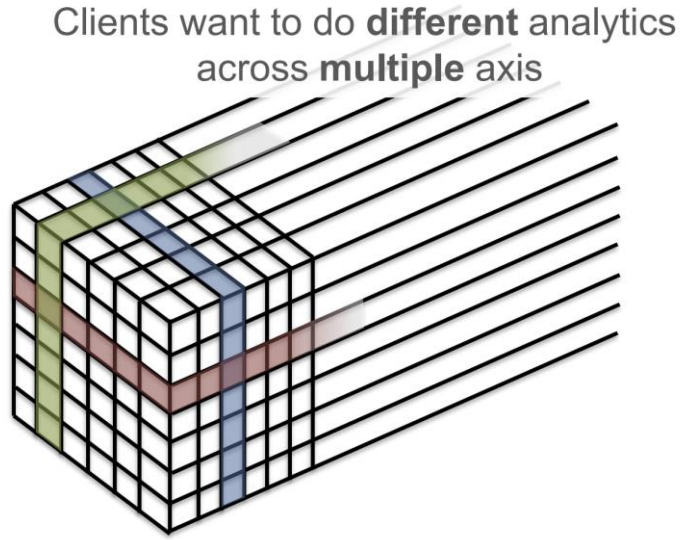


IOR single node performance

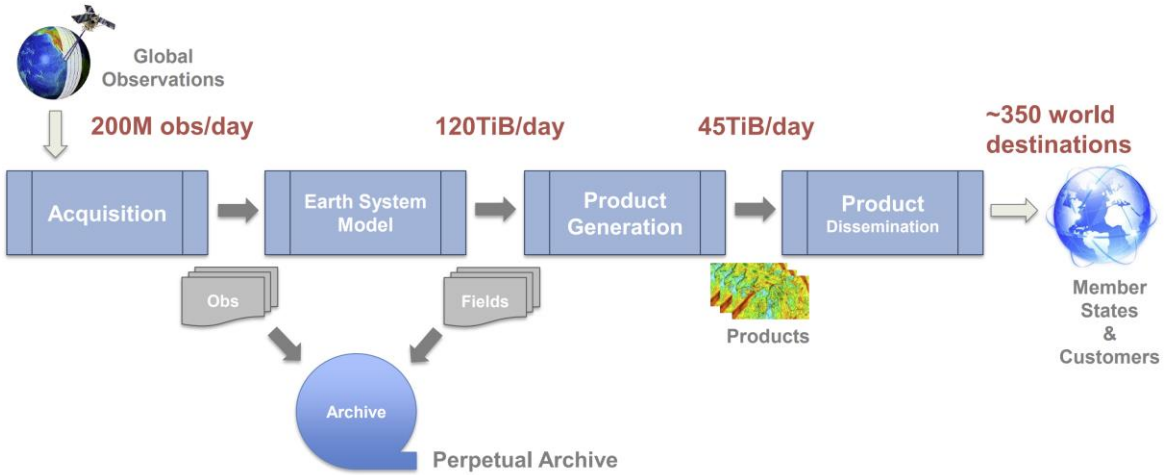




Performance Comparison

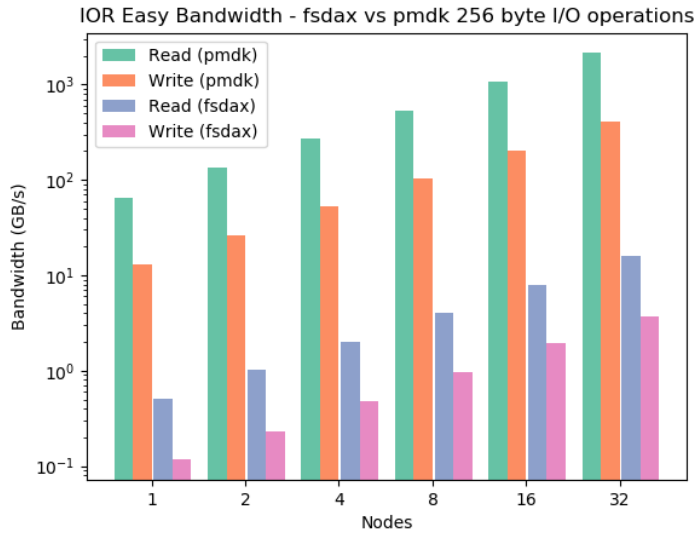


- Pursuing optimal I/O for applications
  - Weather forecasting workflows
- End-to-end workflow performance important
  - Simulation (data generation) only one part
- Consumption workloads different in dimension from production workloads



Hardware setup

- 34 node cluster with 3TB of Intel DCPMM per node
  - 1st generation optane, 256GB DIMMs
- 2 Cascade lake CPUs per node
  - each with 1.5TB of DCPMM and 96GB of DRAM



Hardware performance

- 10 client nodes – 23 filesystem nodes GekkoFS
  - Client nodes are the same as the filesystem nodes
  - Currently using TCP/IP for communication
  - Only a single network rail

```
[RESULT] BW   phase 1            ior_easy_write              66.830 GB/s : time 379.75 seconds
[RESULT] IOPS phase 1         mdtest_easy_write            1714.050 kiops : time 328.88 seconds
[RESULT] BW   phase 2            ior_hard_write               3.345 GB/s : time 303.36 seconds
[RESULT] IOPS phase 2         mdtest_hard_write             192.137 kiops : time 324.38 seconds
[RESULT] IOPS phase 3                      find              612.380 kiops : time 924.33 seconds
[RESULT] BW   phase 3             ior_easy_read              43.421 GB/s : time 584.49 seconds
[RESULT] IOPS phase 4          mdtest_easy_stat            1730.970 kiops : time 325.67 seconds
[RESULT] BW   phase 4             ior_hard_read              22.053 GB/s : time  46.02 seconds
[RESULT] IOPS phase 5          mdtest_hard_stat            1578.850 kiops : time  39.47 seconds
[RESULT] IOPS phase 6        mdtest_easy_delete             876.839 kiops : time 642.89 seconds
[RESULT] IOPS phase 7          mdtest_hard_read             498.590 kiops : time 125.00 seconds
[RESULT] IOPS phase 8        mdtest_hard_delete              46.809 kiops : time 1335.17 seconds
[SCORE] Bandwidth 21.5098 GB/s : IOPS 570.867 kiops : TOTAL 110.812
```

THE UNIVERSITY of EDINBURGH

ECMWF

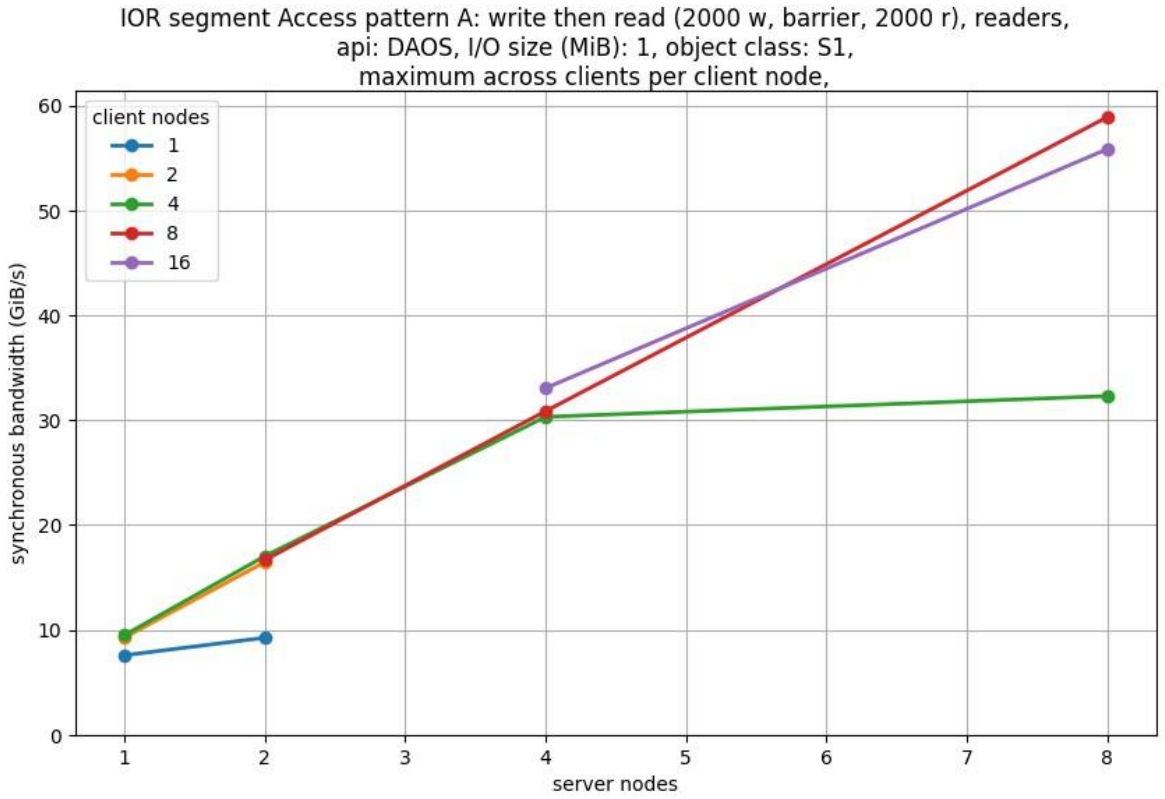
epcc

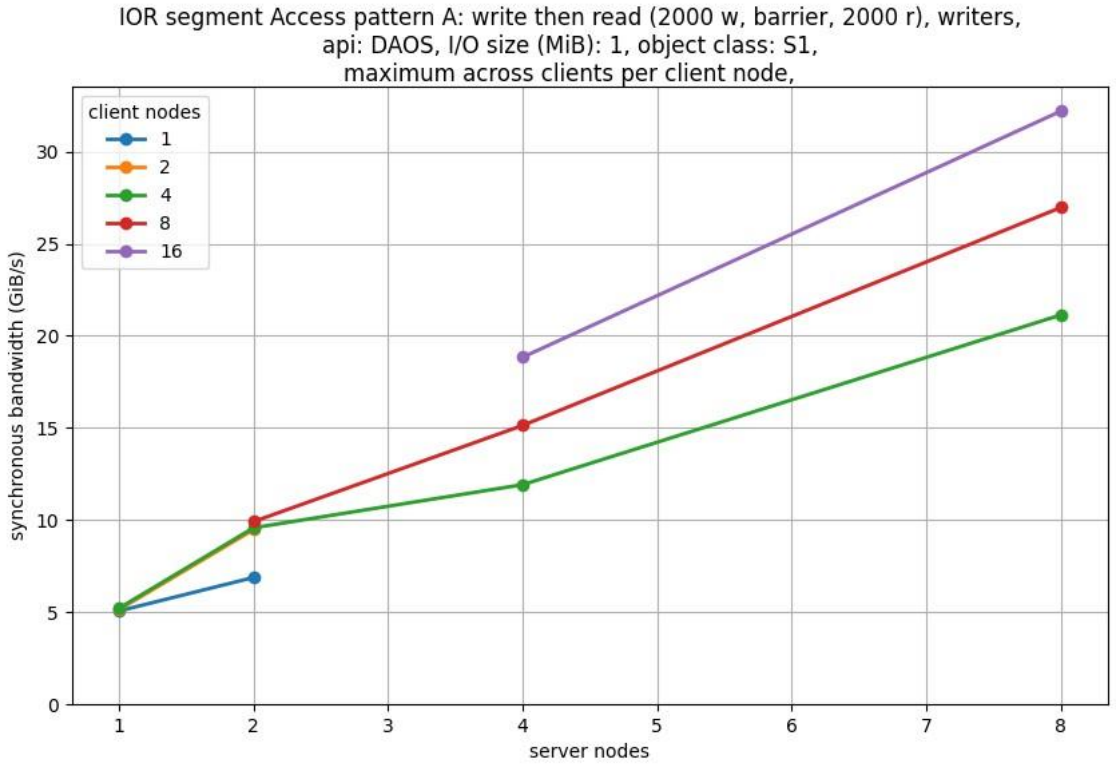
Object store approach

- Data not naturally clustered into “file” wrappers
  - Individual weather fields 1-10MB
- Object store potentially a more natural fit
  - Each weather field is an object
  - Meta data can be attached to uniquely locate them within the overall datasets
- Can object stores
  - Enable high performance I/O?
  - Enable distributed functionality?
  - Enable granular access?
  - Enable production level functionality?

DAOS

- Good bulk I/O performance

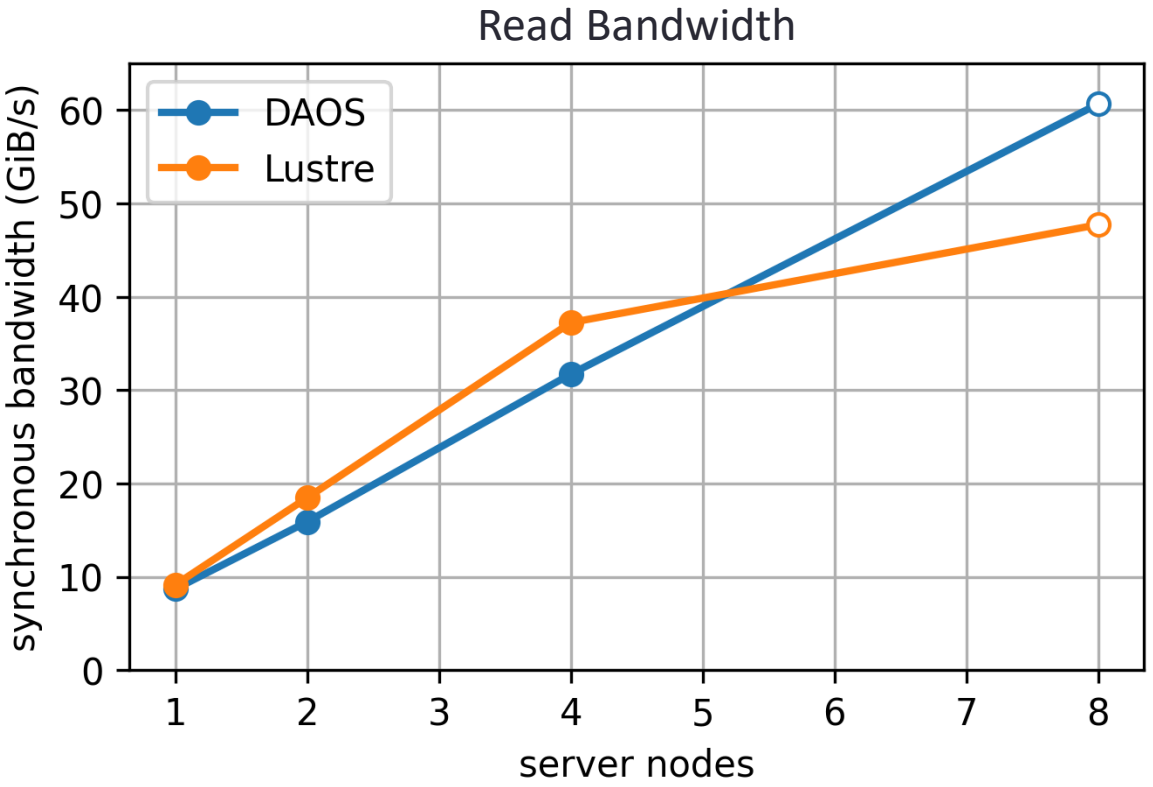


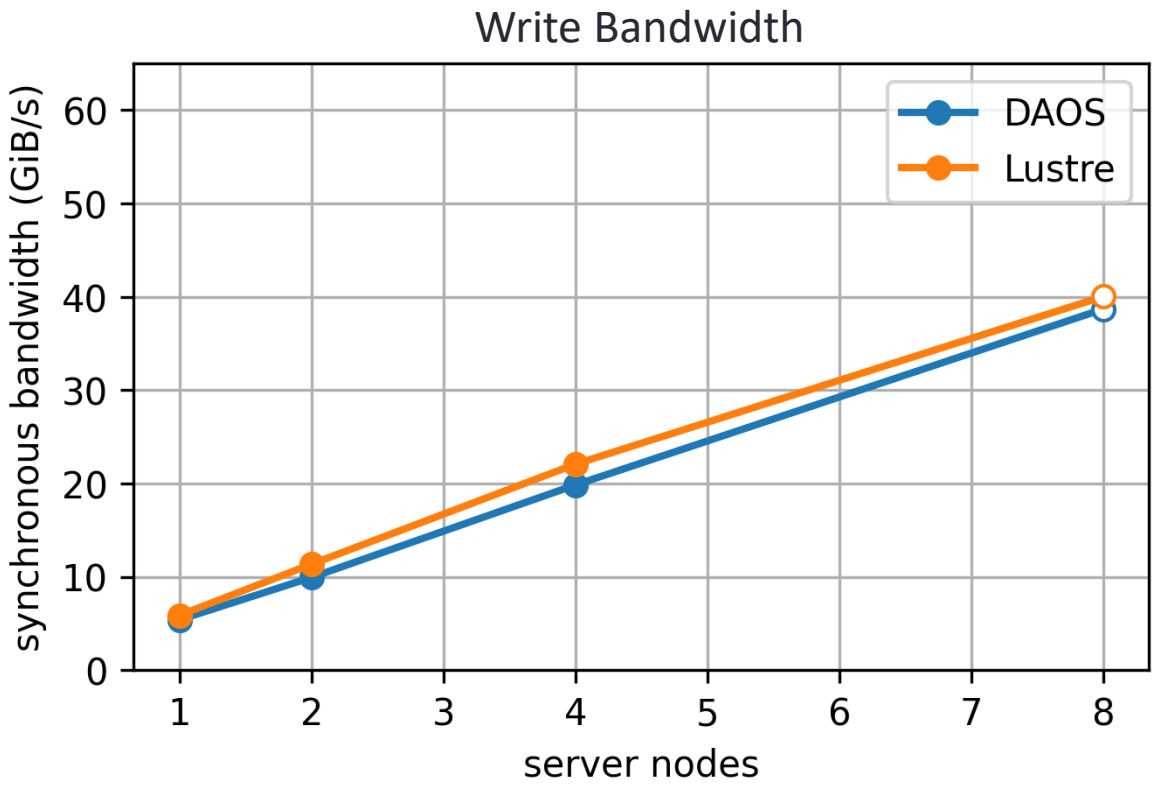


Performance Comparison Hardware configuration

- Setup compute nodes with Optane memory as DAOS server nodes or Lustre server nodes
  - Comparison of Lustre and DAOS on the same hardware
- DAOS server nodes
  - 2 DAOS engines per node (with workers)
  - PMDK/Ext4 filesystem storage backend
- Lustre nodes
  - 1 MDS with 2 targets
  - 2 OSTs per server node
  - Ext4 local storage backend

IOR bulk I/O performance comparison

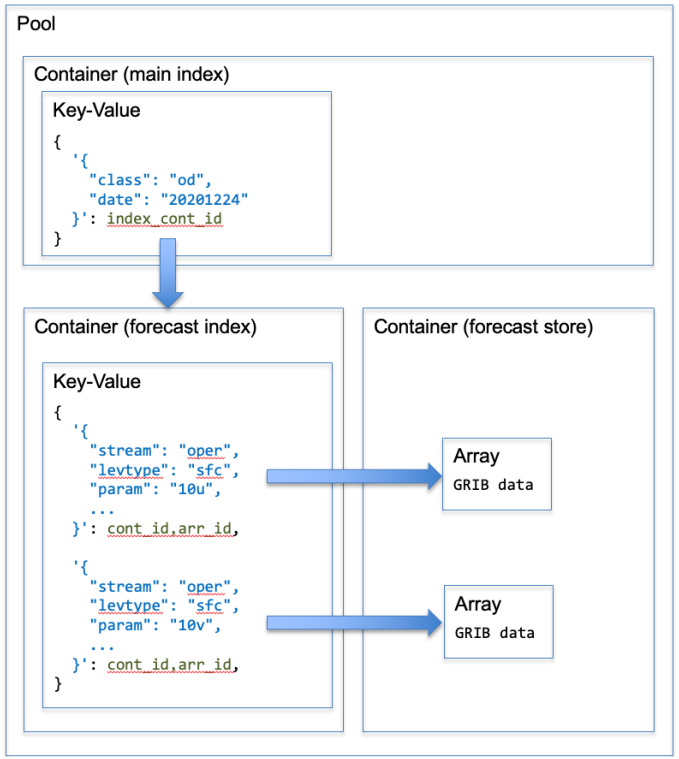




- IOR (easy) benchmark: Segments mode
  - Segments: 100MB (size: 1MB Segment count: 100)
  - POSIX API for Lustre, DAOS API for DAOS

Application like benchmark: Field I/O

- DAOS Field I/O benchmark implements domain-specific object store
  - Indexing with containers and arrays for data storage
- Lustre (POSIX) port of application – object interface
  - Pools are a directory
    - Containers are sub directories within a pool
      - Key-Value objects are sub directories within a container
      - Key is index file
      - Array data separate files
- Two benchmark approaches
  - Pattern A: Separate I/O phases (write then read)
  - Pattern B: Mixed I/O phases (write and read at the same time)

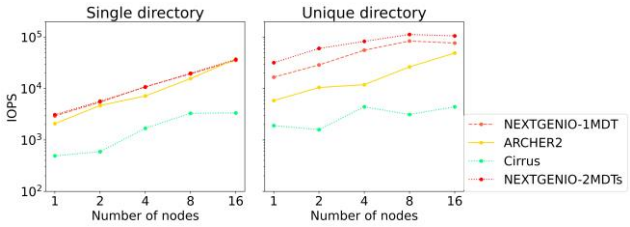


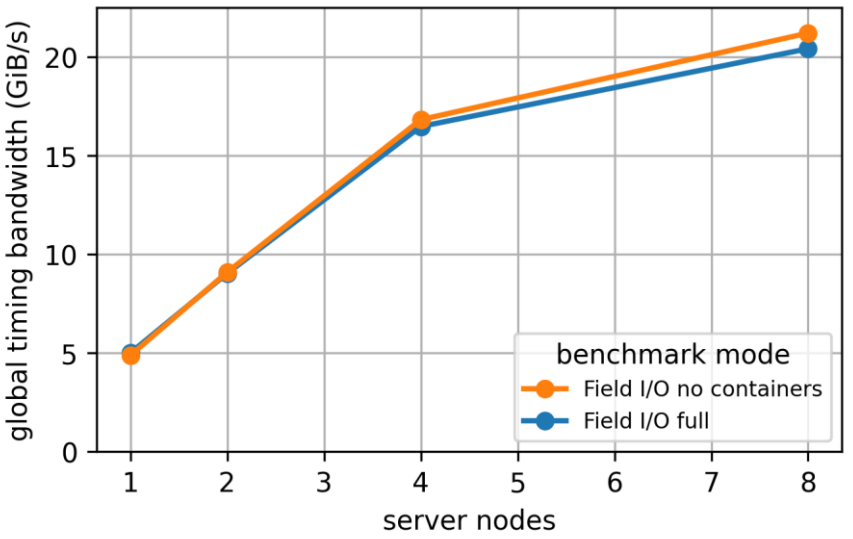
Pattern A: 1MB

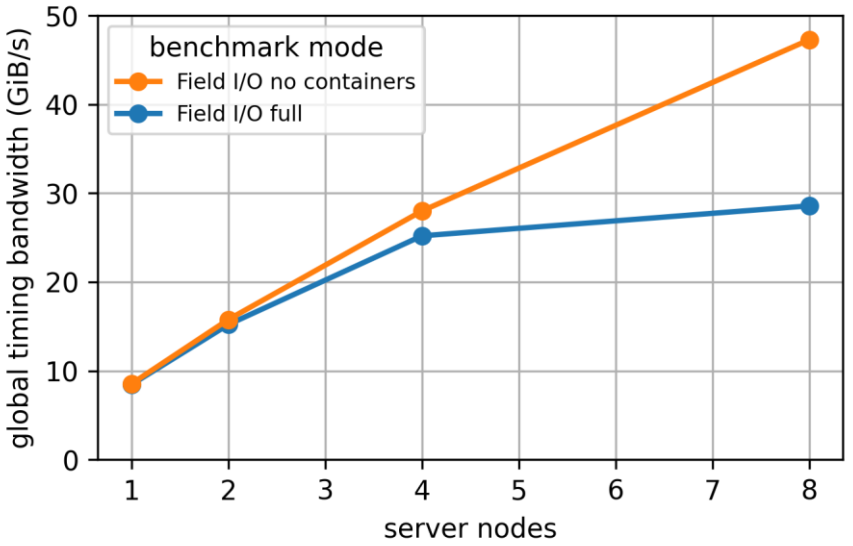
Lustre

DAOS

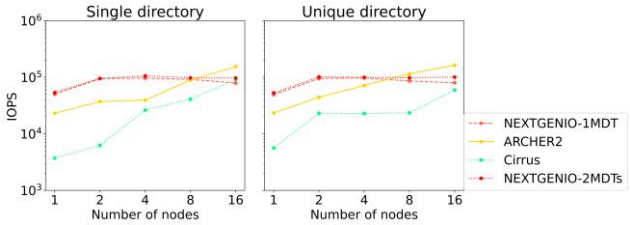
Read

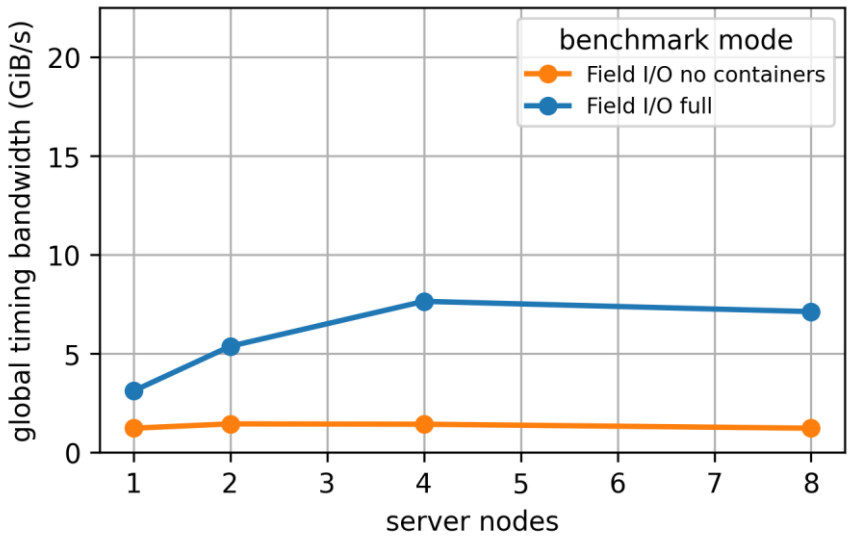


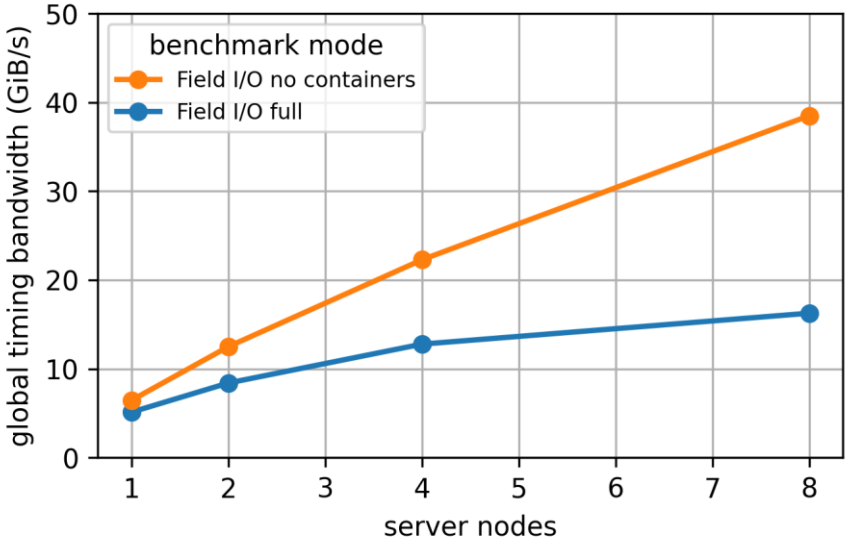




Write





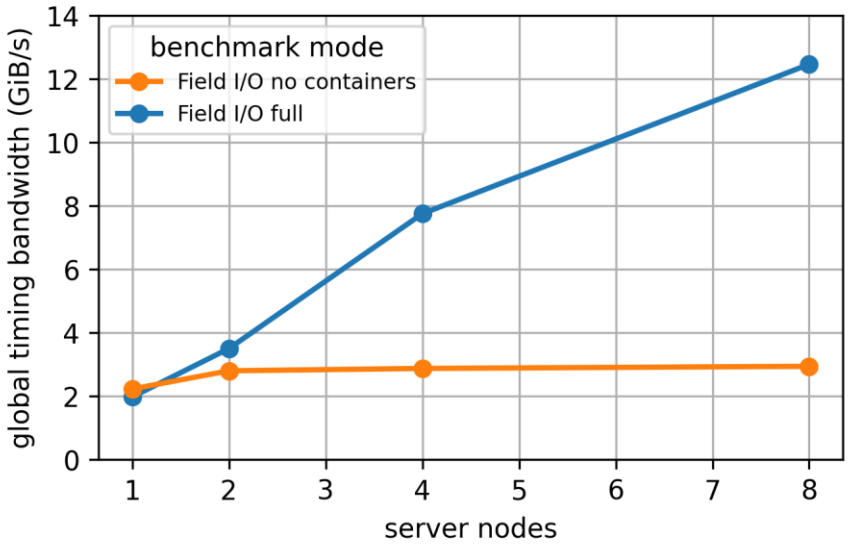


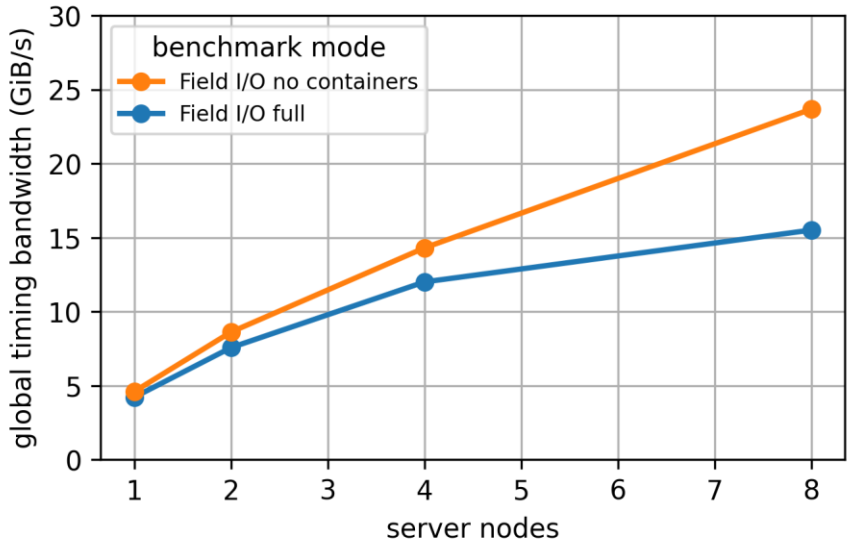
Pattern B: 1MB

Lustre

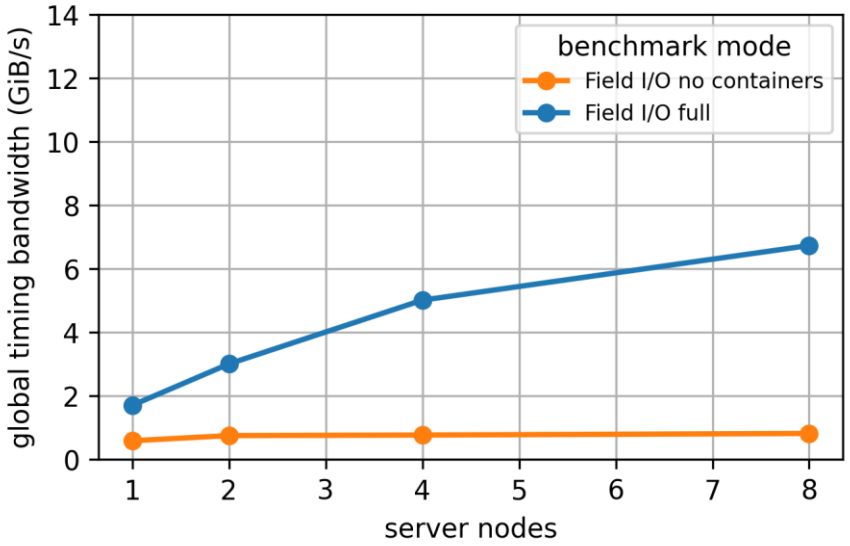
DAOS

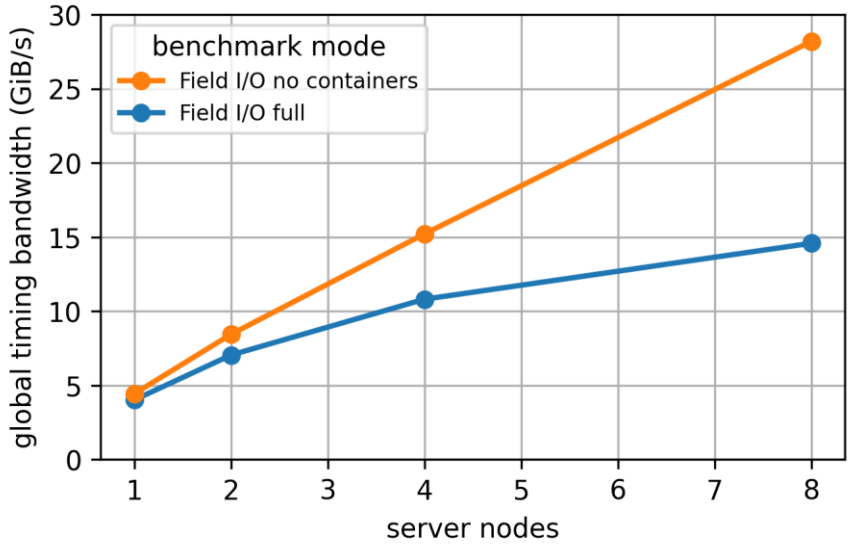
Read





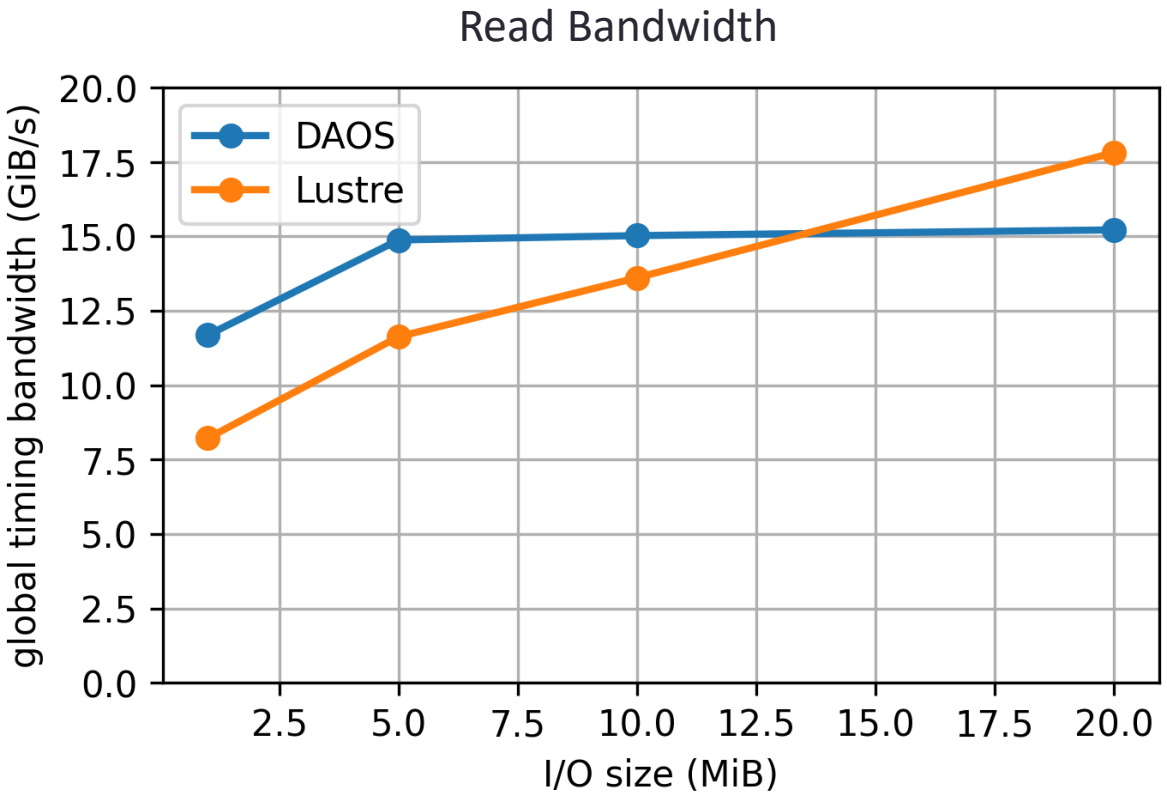
Write

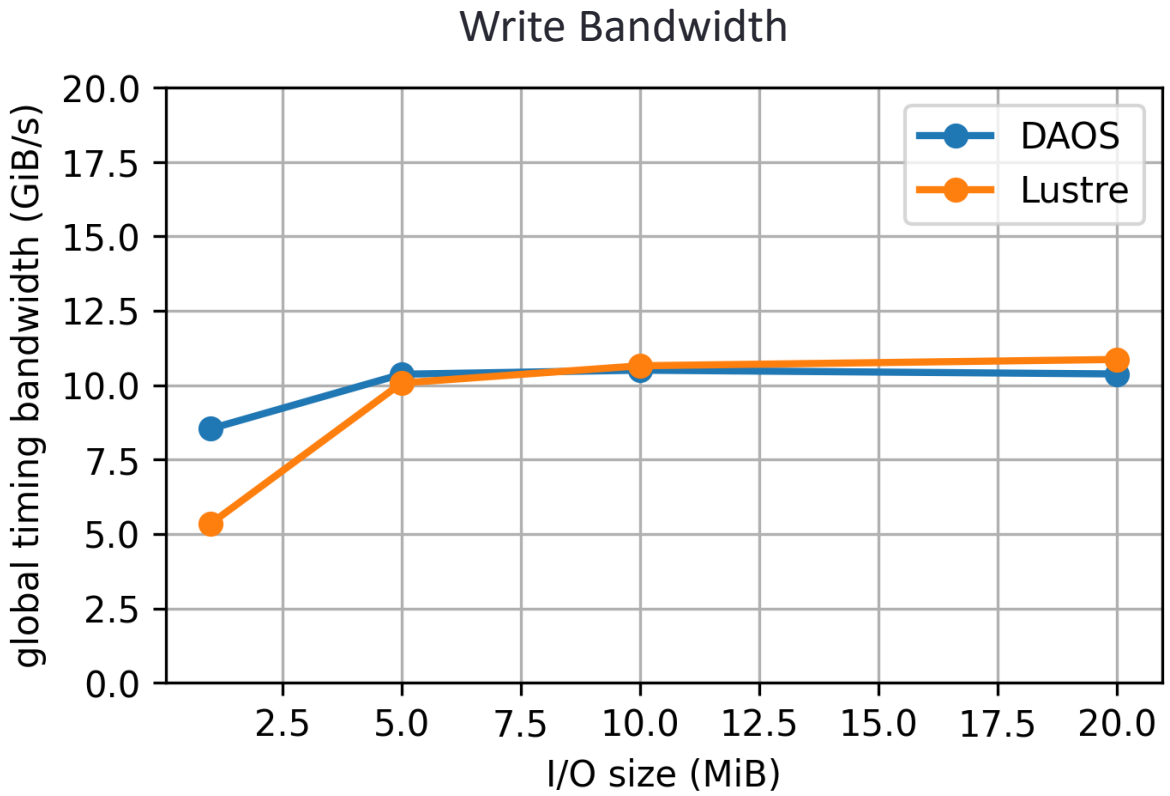




Data size

- Pattern A:
  - 2 server nodes 4 client nodes





Summary

- High performance I/O hardware can support significant performance for applications
  - NVMe as well as NVRAM
- Object store provides more optimised performance for variable I/O patterns
  - Mixed read and write workloads
  - High metadata operations
- Filesystems still provide excellent performance for bulk I/O
- Production functionality can complicate the performance picture
  - Replication/protection
  - DAOS has similar functionality