OAK RIDGE National Laboratory

Tuning I/O Performance on Summit – HDF5 Write Use Case Study

Presented by Bing Xie

Bing Xie, Houjun Tang, Suren Byna, Quincey Koziol, Sarp Oral

ORNL is managed by UT-Battelle, LLC for the US Department of Energy

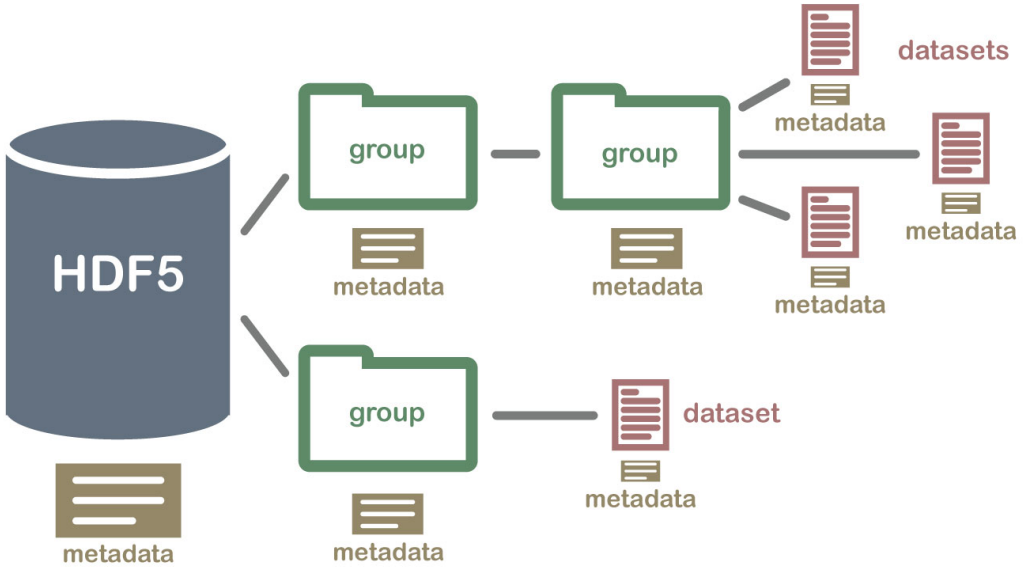
U.S. DEPARTMENT OF ENERGY

Target I/O system

- Summit and Alpine at OLCF/ORNL
  - Fastest supercomputer (2018-2020)
  - 250 PB storage capacity at 2.5 TB/s
  - GPFS filesystem

HDF5: Hierarchical Data Format

- HDF5 I/O Library
  - Widely used in HPC
  - Flexible data format
  - Hierarchical data structure
  - Tunable I/O parameters



HDF5 Write Performance Issue on Summit

- Problem
  - Observed poor write performance with HDF5
- Contribution
  - Configured IOR benchmarks to emulate a plasma physics I/O kernel (VPIC)
  - Studied Darshan eXtended Trace (DXT)
  - Identified a performance issue with mismatching data layout on filesystem
  - Obtained up to 100x write performance improvement in VPIC

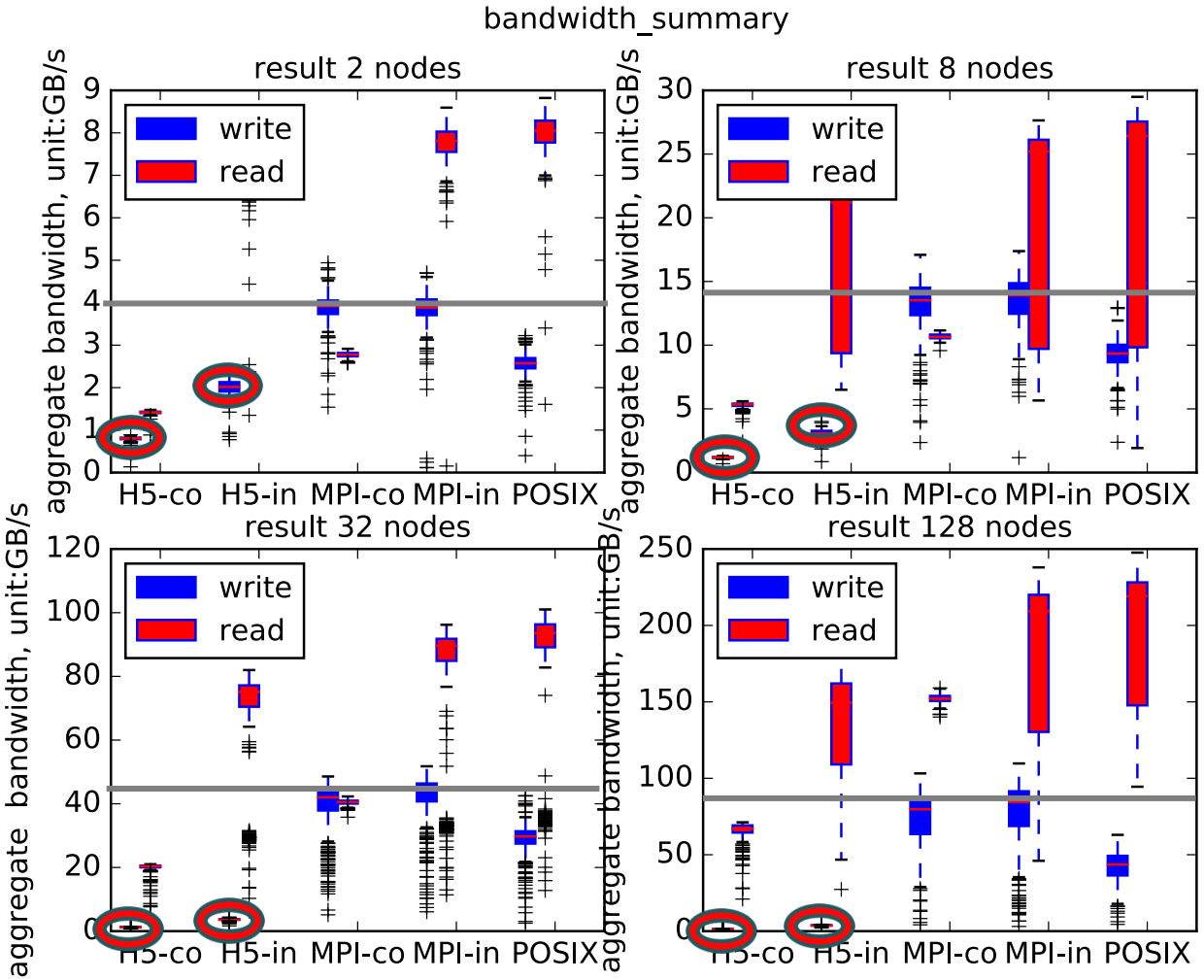
Motivation: Poor write performance of HDF5 on Summit

- IOR configurations (Simulating VPIC I/O behaviors)
  - 256MB write bursts, 32MB block size, all the MPI ranks write to a single file

Aggregate Read/Write Bandwidths
on 2, 8, 32, 128 Nodes

- H5-co: HDF5 collective I/O
- H5-in: HDF5 independent I/O
- MPI-co: MPI collective I/O
- MPI-in: MPI independent I/O
- POSIX: POSIX I/O

1. HDF5 presents poor write performance



Challenges

- High performance variability
- Limited filesystem visibility for end-users
- Tunable performance in HDF5 with many parameters

Our Approach

- Highly variable, but statistically stable
  - **execute IOR benchmarks with POSIX, MPI, and HDF5 with different configurations**
  - **Repeat IOR runs with different compute-node allocations at different times**
- Limited visibility for end users
  - **Build an IOR benchmarking platform and execute the platform with different IO APIs, on different scales, and with different HDF5 configurations.**
  - **Split I/O performance measures to open, read/write, close**
  - **Expert knowledge on tunable space in HDF5**
- Quantitative data analysis
  - **Use Boxplots to present the benchmarking data across scales, APIs, and HDF5 configurations**

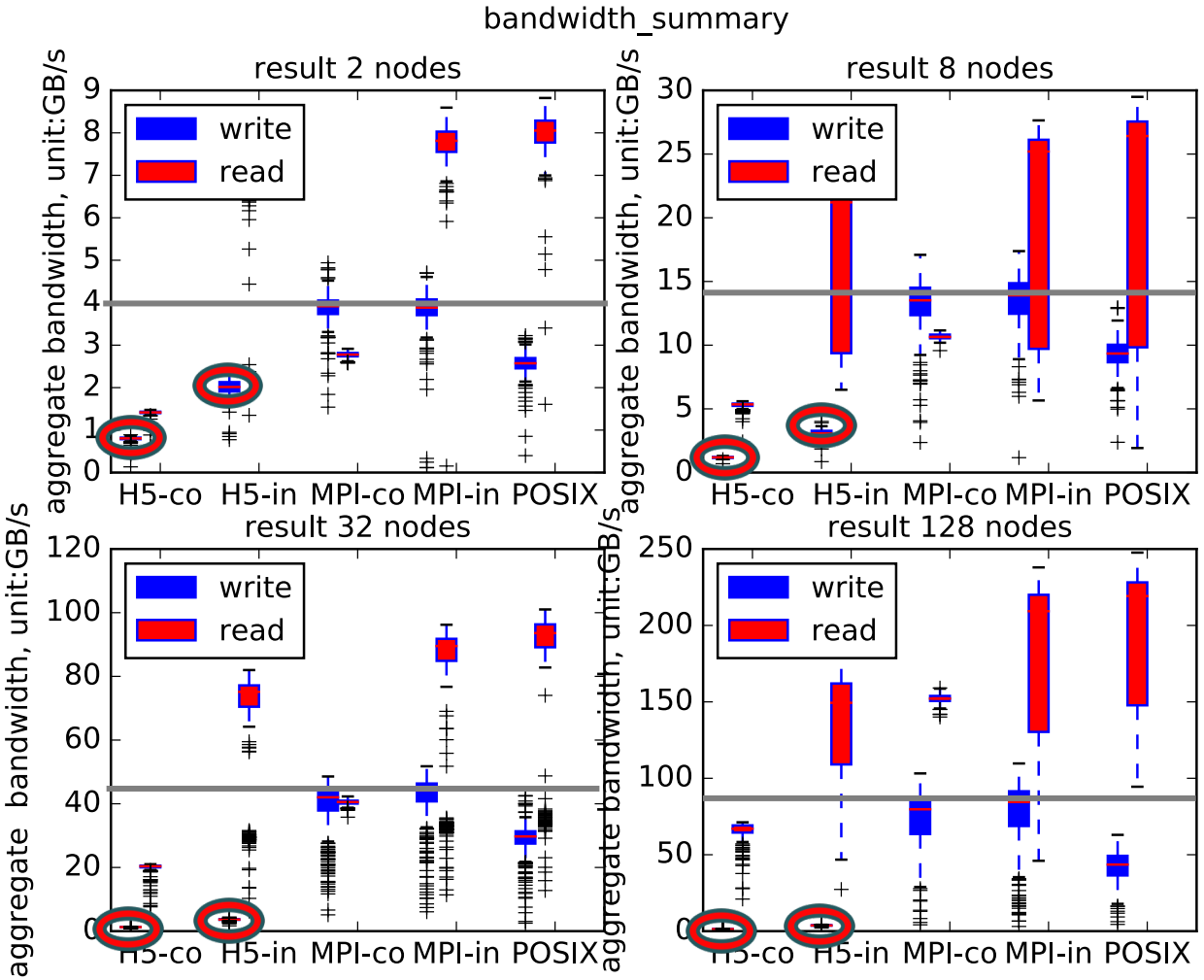
Motivation: Poor write performance of HDF5 on Summit

- IOR configurations (Simulating VPIC I/O behaviors)
  - 256MB write bursts, 32MB block size, all the MPI ranks write to a single file

Aggregate Read/Write Bandwidths
on 2, 8, 32, 128 Nodes

- H5-co: HDF5 collective I/O
- H5-in: HDF5 independent I/O
- MPI-co: MPI collective I/O
- MPI-in: MPI independent I/O
- POSIX: POSIX I/O

1. HDF5 presents poor write performance

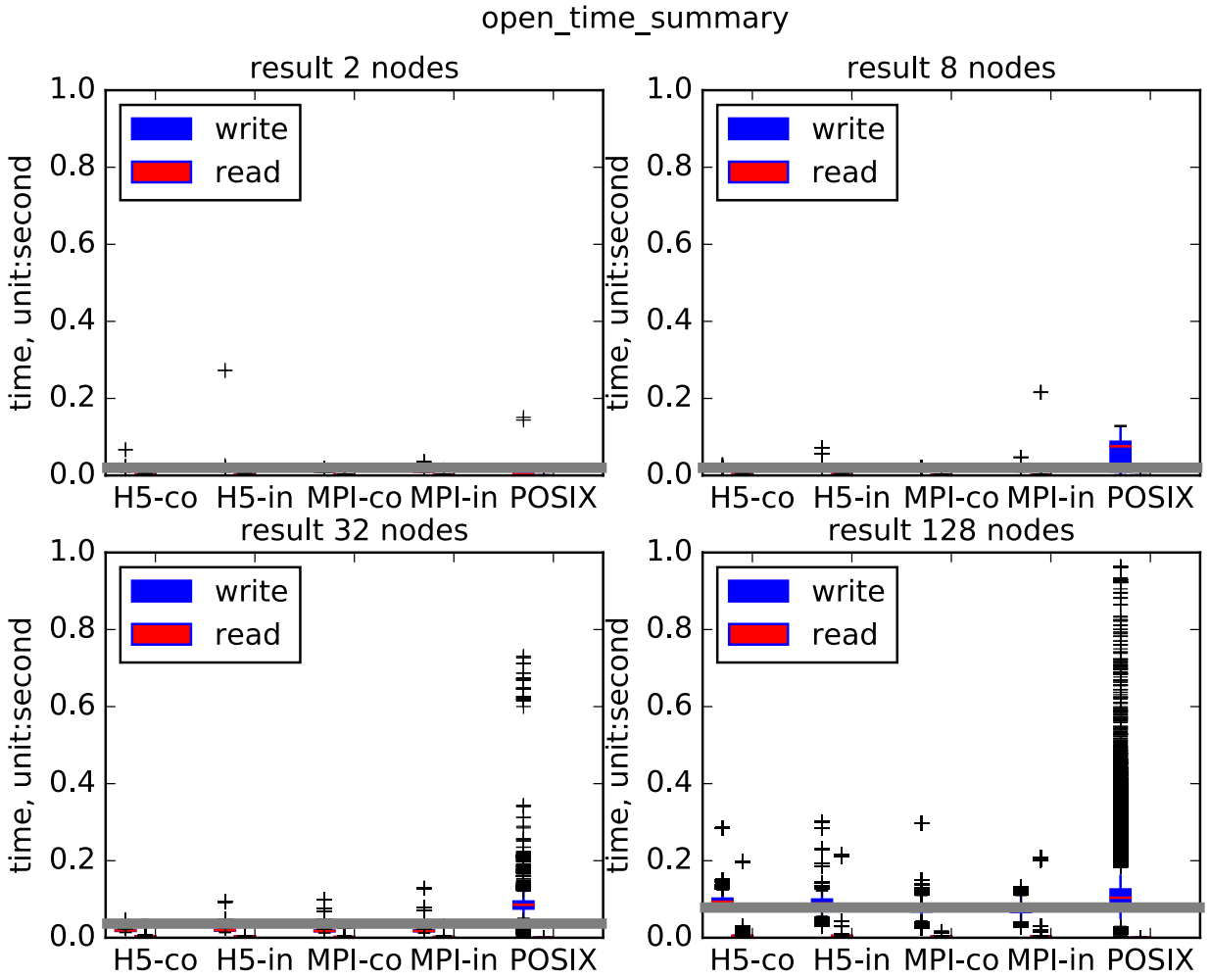


IOR benchmarks on Summit

- IOR configurations (Simulating VPIC I/O behaviors)
  - 256MB write bursts, 32MB block size, all the MPI ranks write to a single file

Open Time on 2, 8, 32, 128 Nodes

- H5-co: HDF5 collective I/O
- H5-in: HDF5 independent I/O
- MPI-co: MPI collective I/O
- MPI-in: MPI independent I/O
- POSIX: POSIX I/O



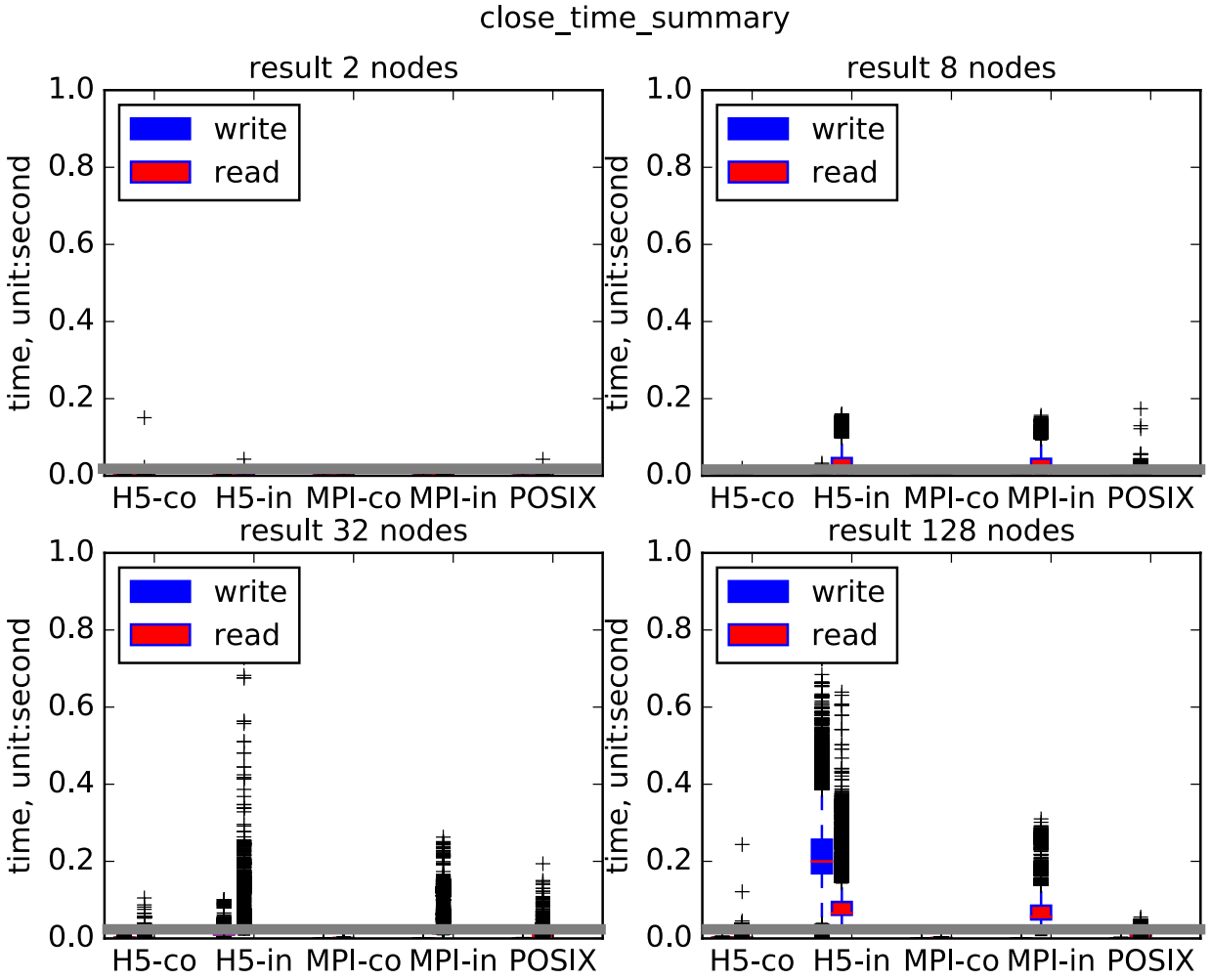
2. HDF5 performs similarly to MPI and POSIX on open times

IOR benchmarks on Summit

- IOR configurations (Simulating VPIC I/O behaviors)
  - 256MB write bursts, 32MB block size, all the MPI ranks write to a single file

Close Time on 2, 8, 32, 128 Nodes

- H5-co: HDF5 collective I/O
- H5-in: HDF5 independent I/O
- MPI-co: MPI collective I/O
- MPI-in: MPI independent I/O
- POSIX: POSIX I/O



3. HDF5 performs similarly to MPI and POSIX on close times

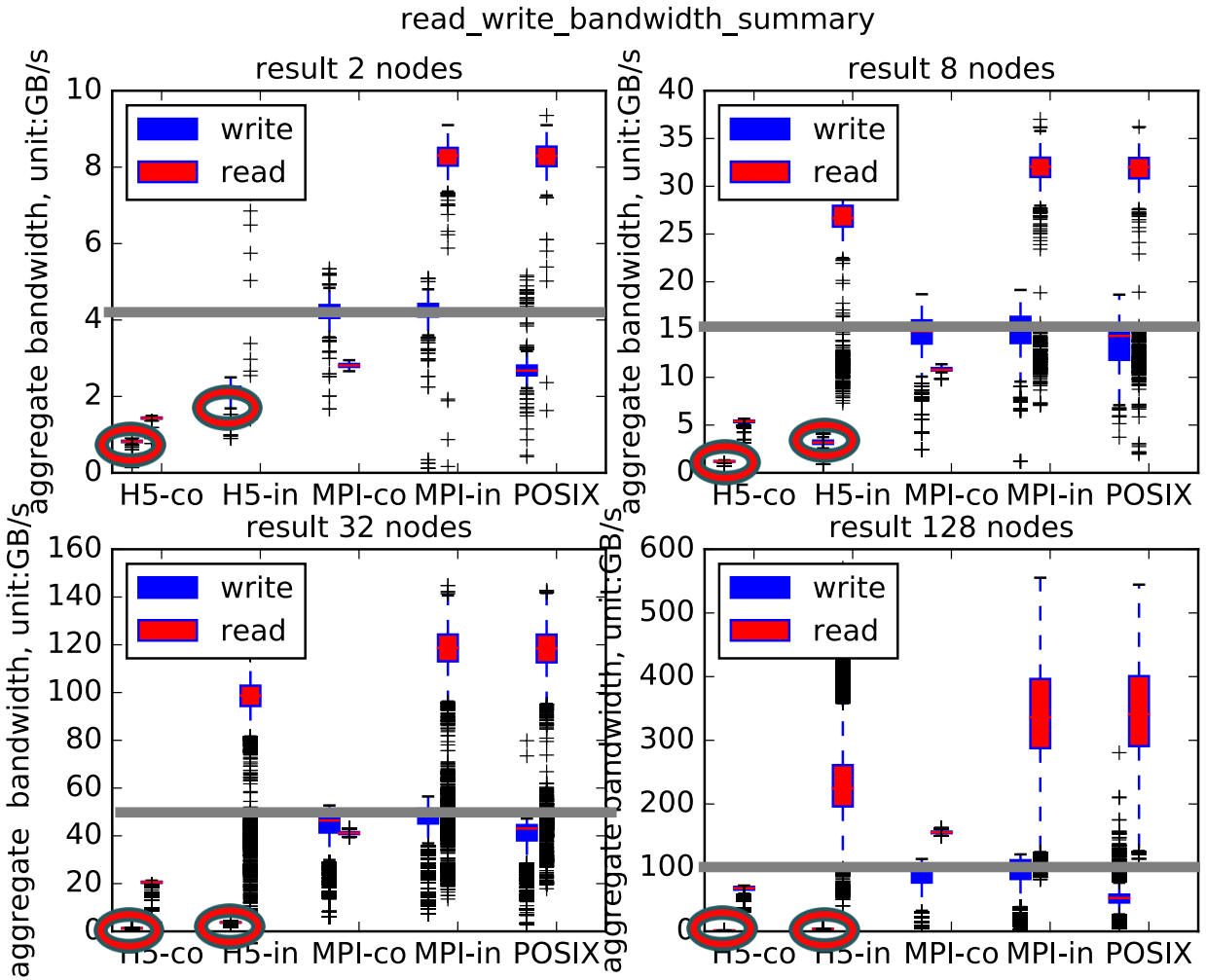
IOR benchmarks on Summit

- IOR configurations (Simulating VPIC I/O behaviors)
  - 256MB write bursts, 32MB block size

Read/Write Bandwidth on 2, 8, 32, 128 Nodes

- H5-co: HDF5 collective I/O
- H5-in: HDF5 independent I/O
- MPI-co: MPI collective I/O
- MPI-in: MPI independent I/O
- POSIX: POSIX I/O

4. HDF5 presents less performance on writes to MPI and POSIX

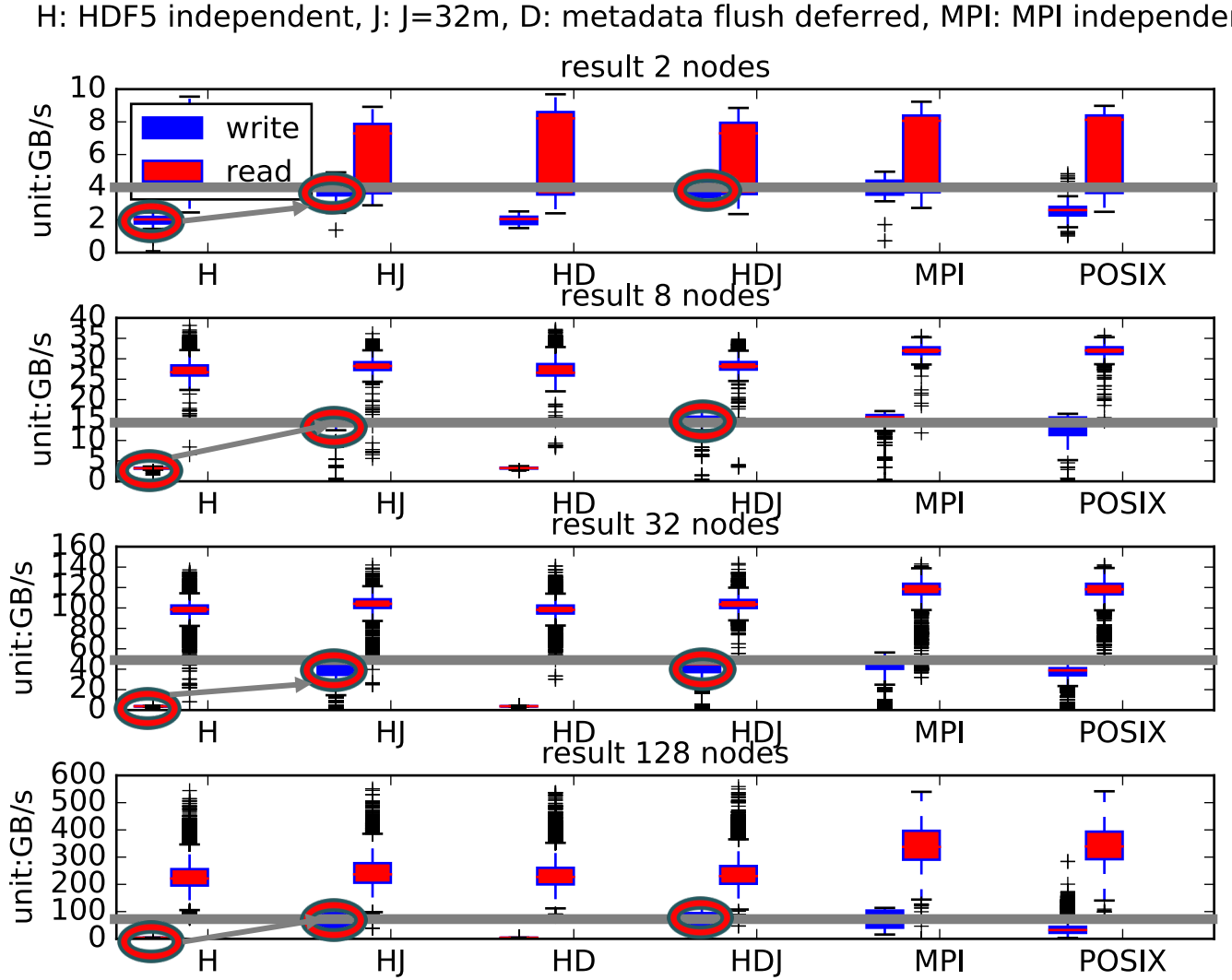


Observations

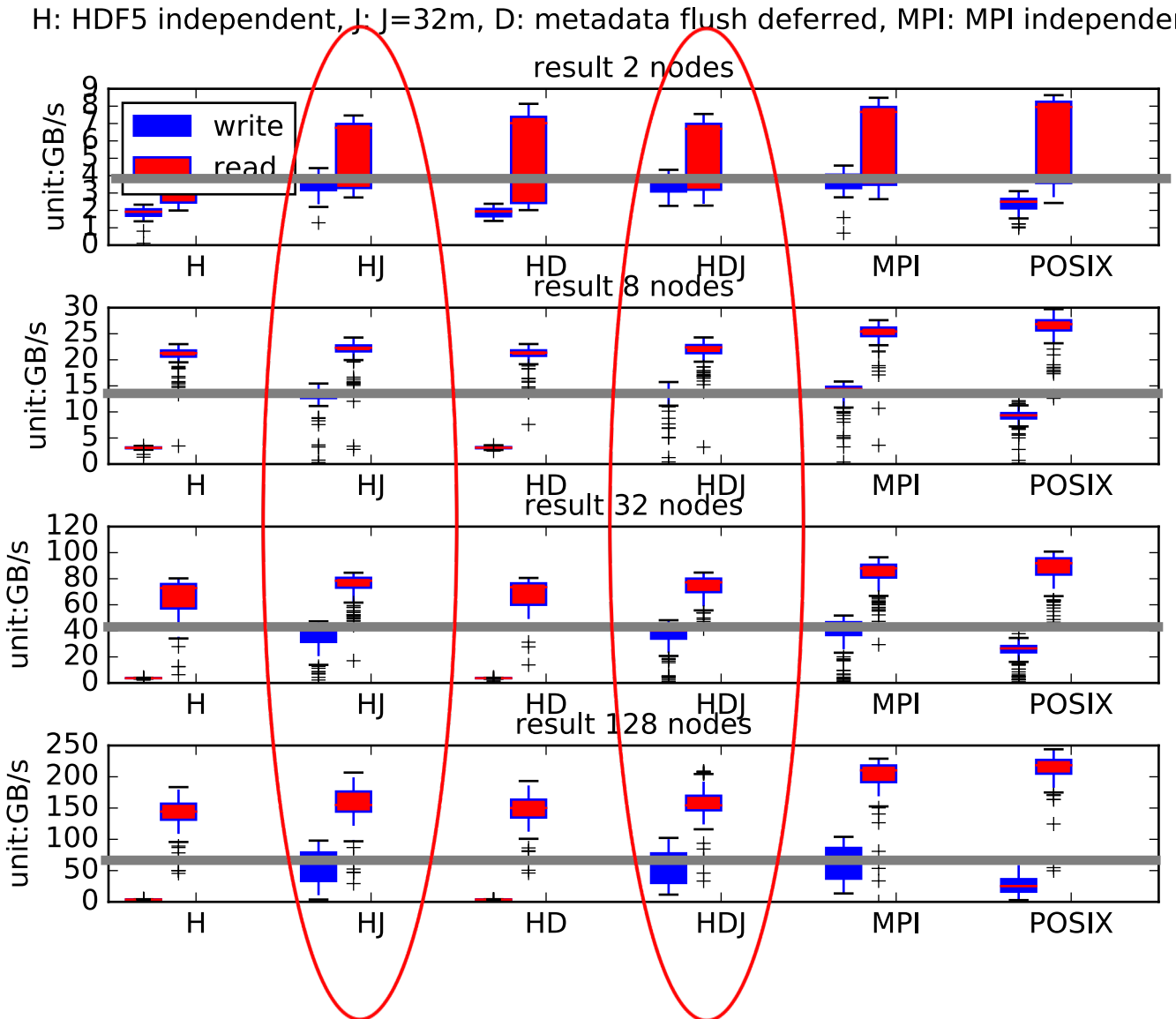
- HDF5 performs similarly to MPI and POSIX on file open and close.
- HDF5 delivers much lower write performance even at write-cache stage (before fsync()).

→ **Expert knowledge: tuning write-cache configurations**

Performance with different write-cache configurations

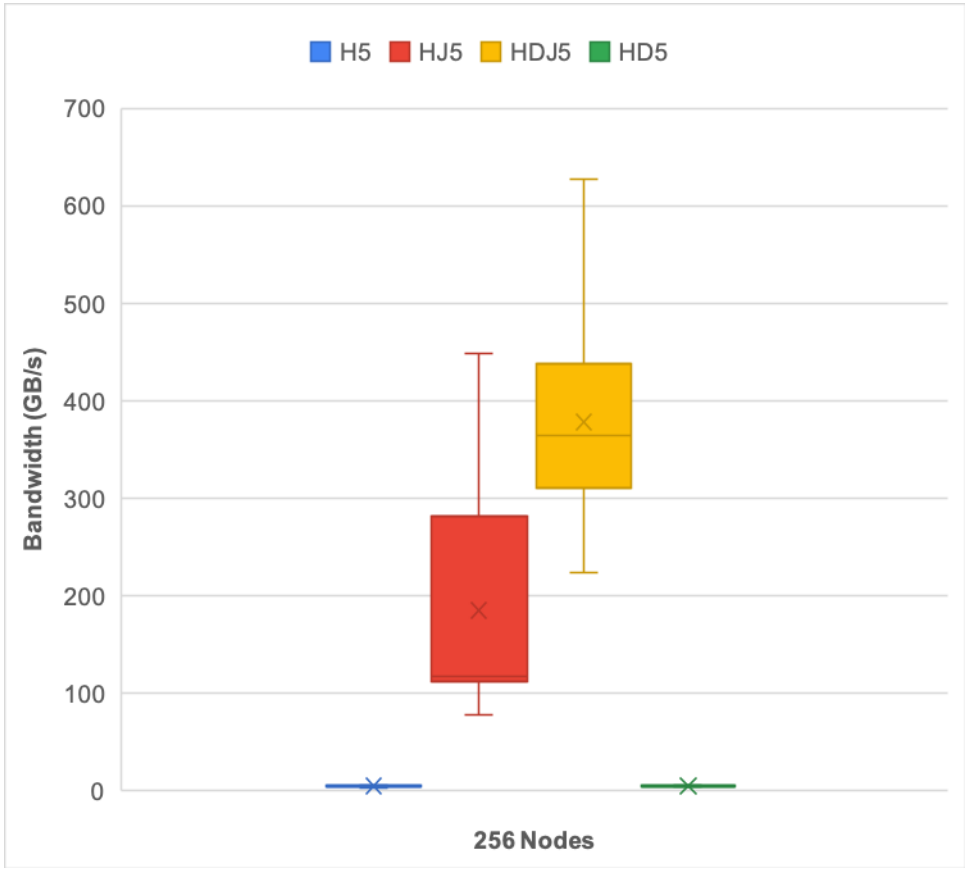


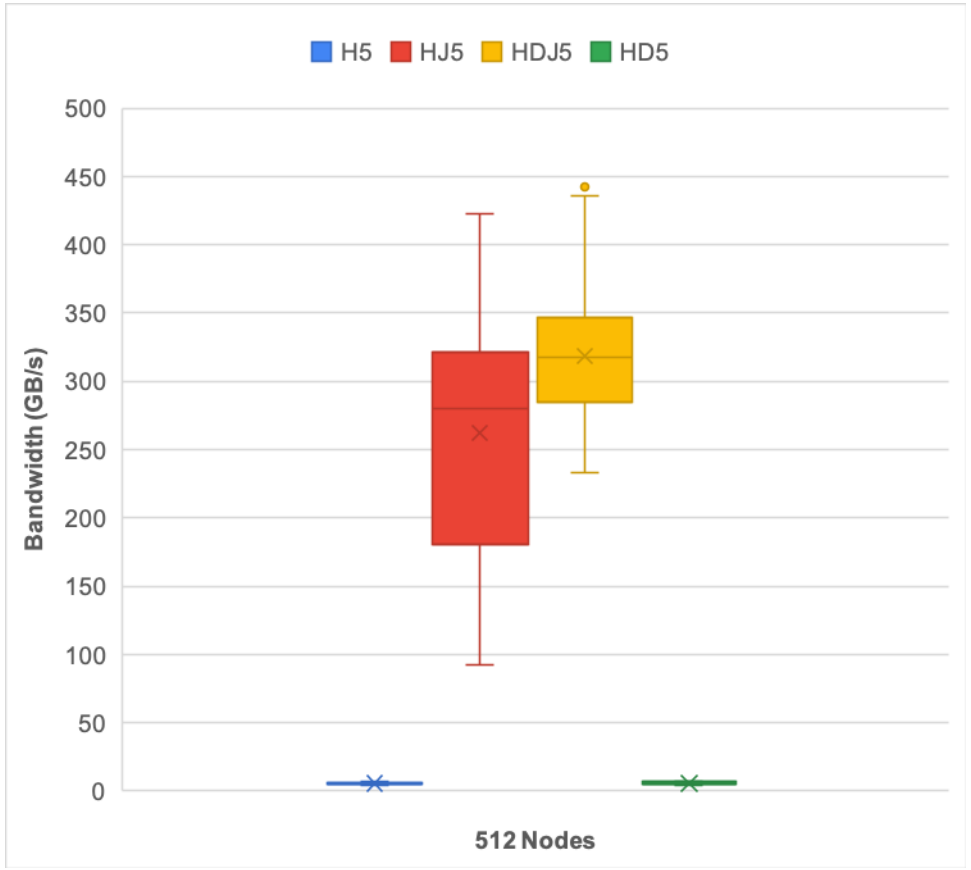
End-to-end Performance with different write-cache configurations



End-to-end Performance with VPIC benchmark

- Up to 100x write performance improvement





Conclusions

- For summit/alpine, setting these alignment by default in HDF5 would be good, but need further evaluation on the impact for various applications and on multiple filesystems.
- Further evaluation is in progress.

Acknowledgements

- This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

- This work is supported in part by the Director, Office of Science, Office of Advanced ScientificComputing Research, of the U.S. Department of Energy underContract No. DE-AC02-05CH11231 and in part by the Exascale Computing Project (17-SC-20-SC), a collaborative effort of the U.S. Department of Energy Office of Science and the National Nuclear Security Administration.