Utilization Trends and I/O Patterns in the Orion-Lustre Filesystem

Authors: Anjus George, Jesse Hanley, Jong Youl Choi, Ahmad Maroof Karimi, Rick Mohr, Christopher Zimmer
National Center for Computational Sciences, ORNL

Presenter: Rick Mohr

Lustre User Group (LUG) 2024

05/07/2024

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

Talk Outline

- Introduction to Orion
  - Multi-tiered Storage Architecture
  - Lustre Specifications on Orion
- Filesystem Characteristics
  - Composition and Utilization
- Prevalent I/O Patterns
  - I/O operations and Data Transfers

The **Orion** Filesystem at OLCF

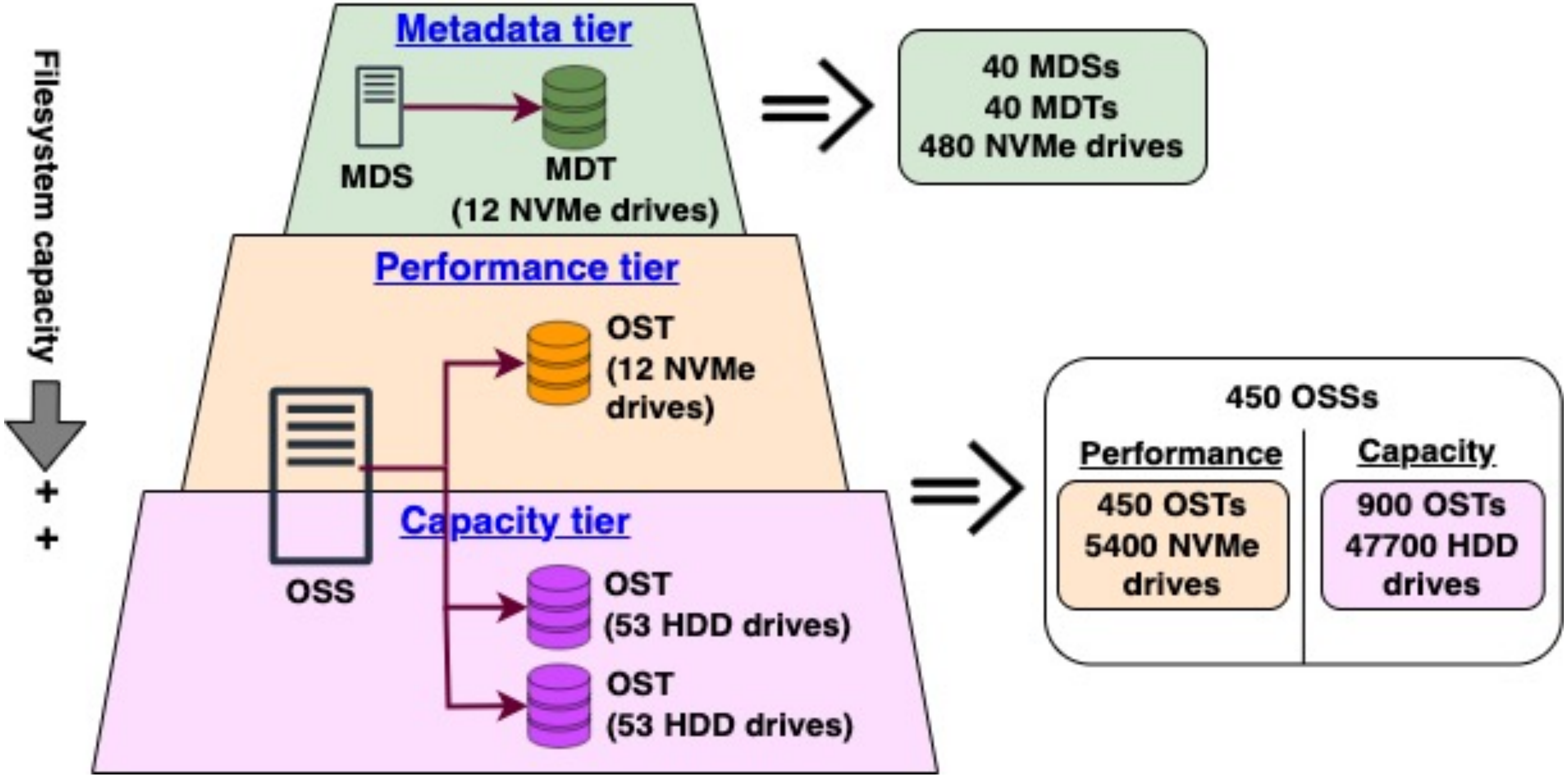
- First US exascale supercomputer Frontier has been in production for over a year.
- The center-wide Lustre filesystem **Orion** serves as primary storage for Frontier.
- Orion uses HPE Cray’s ClusterStor E1000 storage platform.

Multi-tiered Storage Architecture

- Orion implements a multi-tiered storage architecture.
- It comprises 3 tiers unified under a single POSIX namespace.
  - Metadata tier: Stores metadata and first data chunk for all files.
  - Performance tier: Enhances performance for small few MB size files.
  - Capacity tier: Holds majority of data for larger files.

Tier	Capacity	Read BW	Write BW
Metadata	10.0 PB	0.8 TB/s	0.4 TB/s
Performance	11.5 PB	10.0 TB/s	10.0 TB/s
Capacity	679.0 PB	5.5 TB/s	4.6 TB/s

Multi-tiered Storage Architecture (cont’d)



Backend Filesystem

- Backend file system for MDTs/OSTs is ZFS v2.1.7
- Redundancy is handled using ZFS dRAID

Target	# drives	Data blocks	Parity blocks	Spares
MDT (NVMe-based)	12	9	2	1
OST (NVMe-based)	12	9	2	1
OST (HDD-based)	53	11	2	2

Lustre Specifications on Orion

- Lustre version 2.15 w/ vendor patches
- Utilize two OST pools for file placement
  - "performance" for all NVMe OSTs
  - "capacity" for all HDD OSTs
- Distributed Namespace (DNE) used to spread project directories across all MDTs
  - Only utilizing remote directories (DNE1) at this point
  - No striped directories (yet)
- File layouts take advantage of Data on MDT (DoM), Self Extending Layouts (SEL), and Progressive File Layouts (PFL)

Default File Layout

- PFL to accommodate a wide range of I/O patterns.
- Non-DoM components include an SEL extension.

Tier	Component Length	Stripe Size	Stripe Count	Extension Size	Expected File %
Metadata	256 KiB	256 KiB	1	N/A	70%
Performance	8 MiB	1 MiB	1	64 MiB	18%
Capacity	128 GiB	1 MiB	1	16 GiB	11%
Capacity	Infinity	1 MiB	8	256 GiB	1%

Filesystem Composition

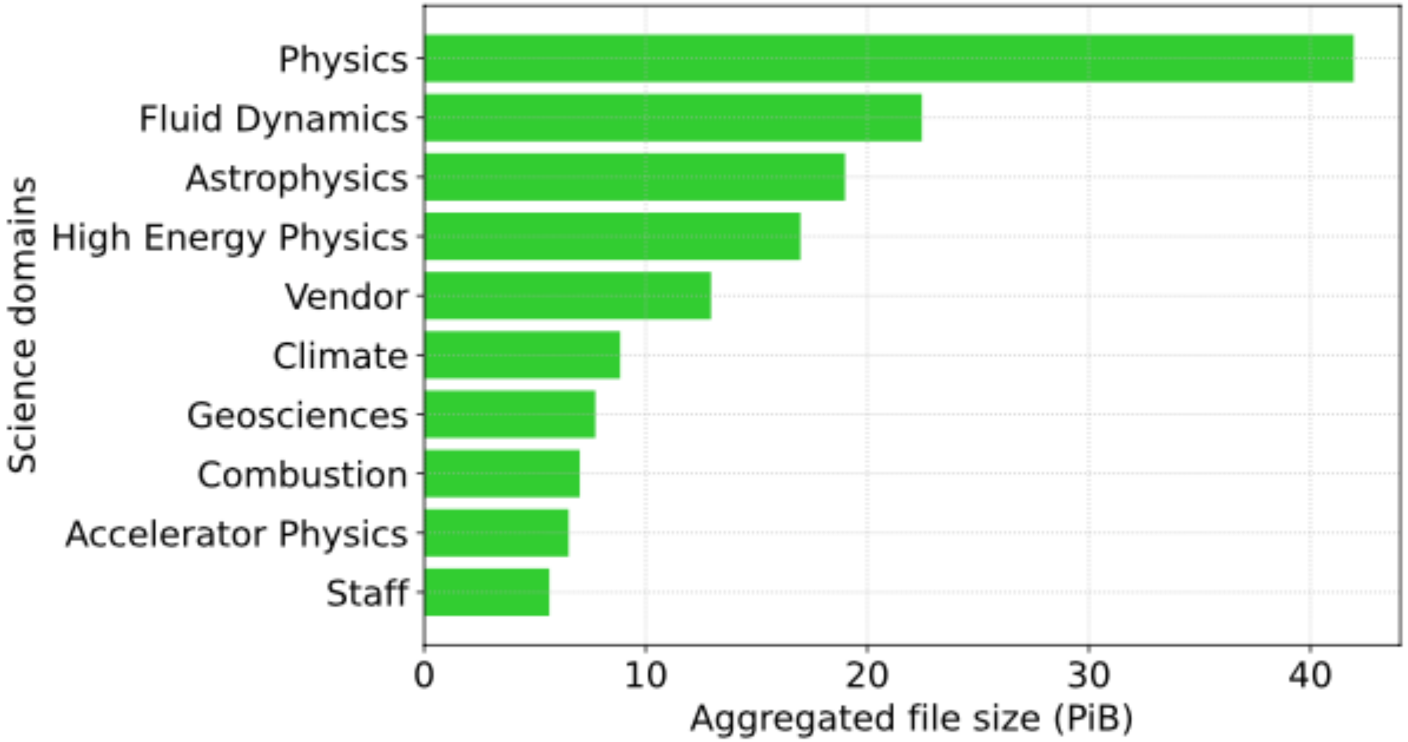
- Profiled Orion using `fprof`[1] profiling tool
- Fprof – lightweight profiler that traverses PFSs
- Table shows a high-level snapshot of Orion.
- Total utilization – 31.12% of the total PFS capacity

Total number of files	2,972,695,978
Total number of directories	102,319,055
Total number of symlinks	16,111,171
Average file size	69.95 MiB
Avg number of entries per directory	29
Aggregated file size	193.65 PiB

(Profiling time frame: mid-March 2024)

Filesystem Utilization

- Conducted profiling on a per-project basis to better understand utilization.
- Profiling results indicated Orion encompassed 488 project directories.
- These directories were mapped to their respective science domains.
  - Physics – 40 PiB, 22.63% of total utilization
  - Fluid Dynamics – 22.46 PiB
  - Astrophysics – 18.99 PiB



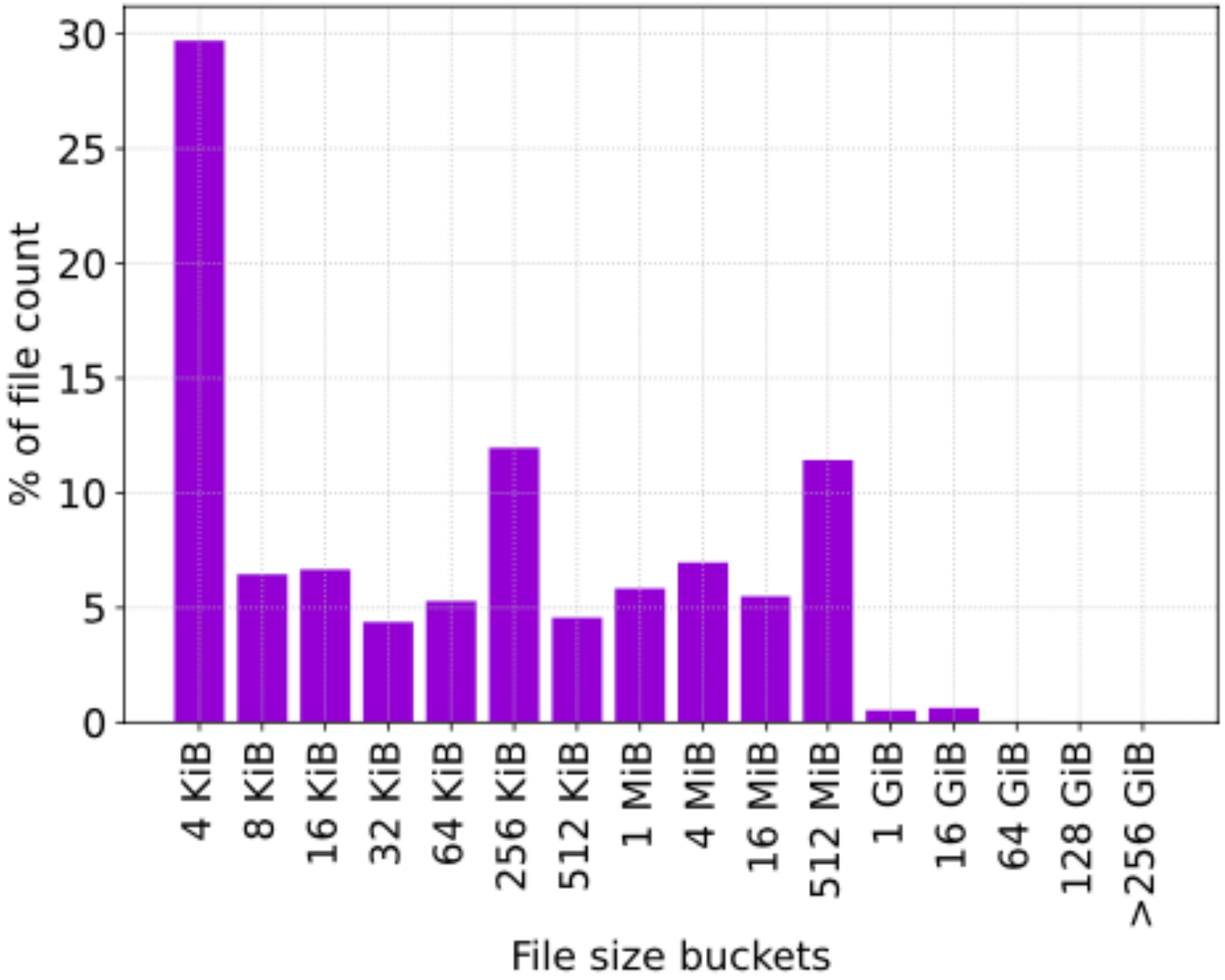
File Size Distribution

- Figure shows the distribution of files in various file size buckets.

- 30% of the files are under 4 KiB, making it the largest file size bucket.

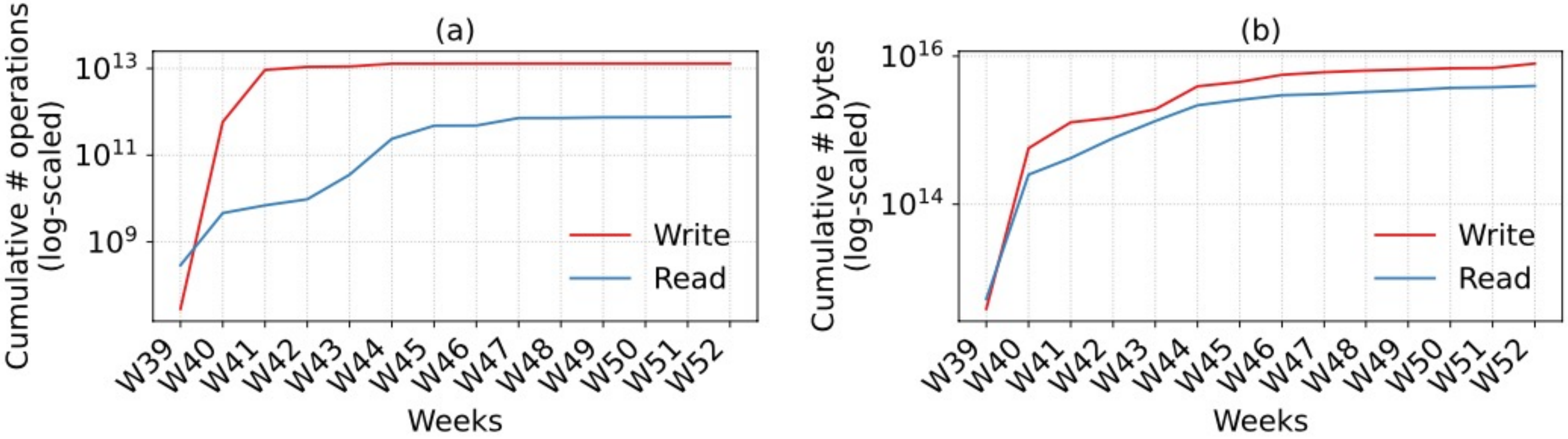
- Next common file size ranges: 64 KiB to 256 KiB and 16 MiB to 512 MiB
  - Each constitute 11% of the total count

- Files larger that 256 GiB are scarce, around 90K files



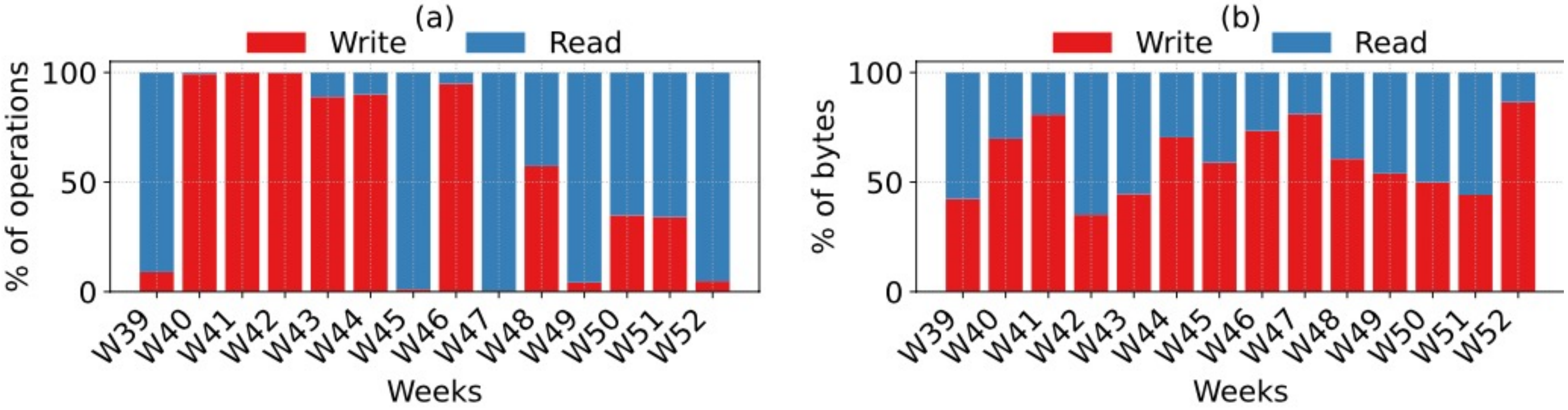
Scale of I/O Operations

- Analyzed logs from Darshan[2] from Oct 2023 to Dec 2023.
- Cumulative of all I/O interfaces including STDIO, MPI-IO, and POSIX
- By 3rd week, the write operations reached 10 trillion (10^{13}).
- The read operations exhibit consistent growth and reached 1 trillion (10^{12}).
- Over a 3-month period, the write transfer size is 10 PB.
- The read transfer size is 1 PB.



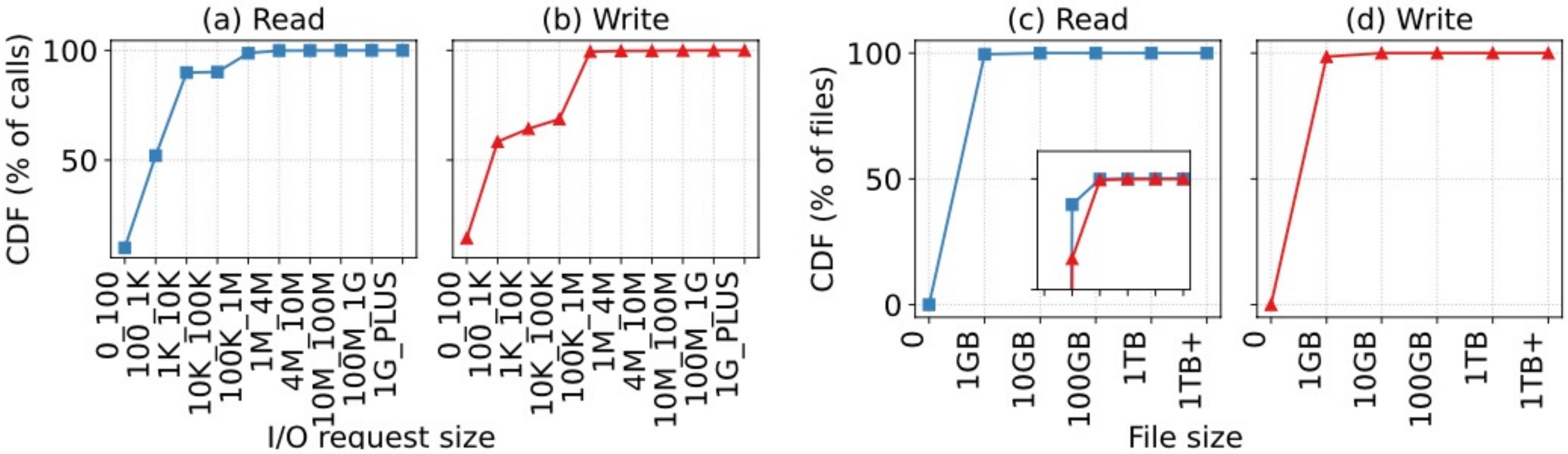
Read/Write Call Distribution

- Figure shows the % of read and write operations in different weeks.
- Difficult to observe a fixed trend for read/write operations.
- Certain weeks show dominance of either read or write.
- # bytes transferred writes dominate read for except 4 weeks.



Dominant Data Transfer Patterns

- Also analyzed the I/O request size and file size used by workloads.
- Write request size is larger than read request size.
- Majority of the files for read and write are < 1GB.
- Zoom in plot shows there are significant write files up to 10 GB.



Summary

- Orion has been operational for over a year catering I/O needs for data intensive HPC applications.
- It is composed of over 2.9 billion files and 102 million directories.
- Current utilization is 31% of the total PFS capacity.
- Over 30% of files are under 4 KiB and files > 256 GiB are scarce.
- Read/Write calls are mostly balanced with write dominating at some instances.
- Weekly dominance of read/write calls indicate large scale jobs running on majority of Frontier nodes.
- Majority of files for read/write are < 1 GB.

Acknowledgements

This research used resources of the Oak Ridge Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC05-00OR22725.

Questions?

References

1. Feiyi Wang, Hyogi Sim, Cameron Harr, and Sarp Oral. 2017. Diving into petascale production file systems through large scale profiling and analysis. In Proceedings of the 2nd Joint International Workshop on Parallel Data Storage & Data Intensive Scalable Computing Systems (Denver, Colorado) (PDSW-DISCS ’17). Association for Computing Machinery, New York, NY, USA, 37–42. https://doi.org/10.1145/3149393.3149399

2. Philip Carns, Kevin Harms, William Allcock, Charles Bacon, Samuel Lang, Robert Latham, and Robert Ross. 2011. Understanding and Improving Computational Science Storage Access through Continuous Characterization. ACM Transactions on Storage (TOS) 7, 3, Article 8 (Oct. 2011), 26 pages. https://doi.org/10.1145/2027066.2027068