

Deep Dive into OLCF Storage Systems ATPESC 2022 - Track 7 - I/O August 10, 2023

Michael J. Brim, Senior R&D Staff

National Center for Computational Sciences (NCCS) Oak Ridge Leadership Computing Facility (OLCF)

Oak Ridge National Laboratory

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



Overview and Goals for this Lecture

- 1. Quick background on NCCS and OLCF approach to HPC storage
- 2. High-level overview of OLCF's center-wide shared storage system
- 3. Deep dive on Lustre and Orion





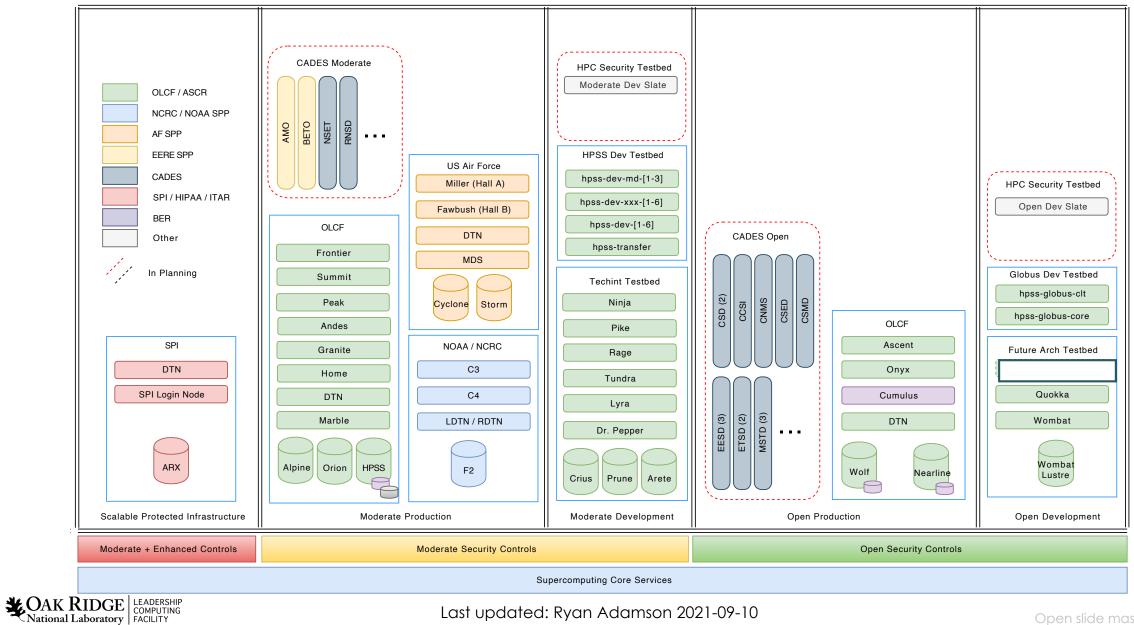
HPC Storage @ NCCS and OLCF

- NCCS organizational overview
- HPC Storage Strategy
- Scratch and Archive Systems



National Center for Computational Sciences

Jpdated:



NCCS HPC Storage Strategy

		OLCF	BER	AirForce	NOAA - NCRC
er ->	Job-term (24 hours or less)	Summit Frontier			
Faste	Short-term (less than 90 days)	Alpine/Orion Arx Wolf	Wolf	Storm Cyclone	F2
	Medium-Term (90 days to 1-3 years)			N/A	N/A
Slower	Long-term (90 days to 20 years)	HPSS/The	emis	N/A	N/A
<- SI	Forever-term (keep data forever)			N/A	N/A



Production Scratch Filesystems



Arx OLCF Mod-enh/GPFS

- 3.3 PB
- 36 GB/s r/w



F2

- NCRC/Lustre-2.12
- 40 PB
- 45 GB/s r/w



Alpine

- Moderate/GPFS
- 250 PB
- 2.5 TB/s r/w



COAK RIDGE National Laboratory

Storm/Cyclone

- AFW/Lustre-2.12
- 2x 7.5 PB
- 45 GB/s r/w
- High resiliency

Wolf

- OLCF Open/GPFS
- 7.7 PB
- 90 GB/s r/w



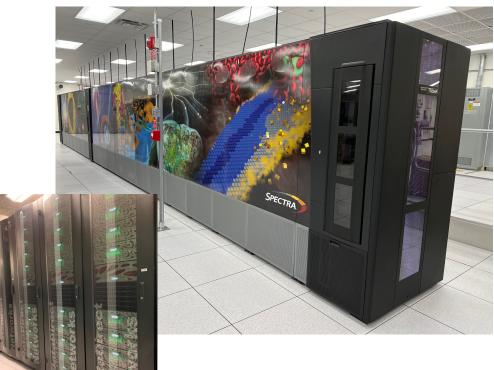
Production Archive Filesystems

HPSS

• HPSS-7.2

COAK RIDGE National Laboratory

- 160 PB of tape (RAIT3+1p)
- 22 PB of disk cache
- 12 GB/s performance



Themis

- IBM Spectrum Archive
- 60 PB of tape (2-way replication)
- 10.2 PB of disk
- 70 GB/s disk performance





The OLCF Center-wide Shared File System

- Why center-wide?
- Spider Architecture and History
- Production File Systems

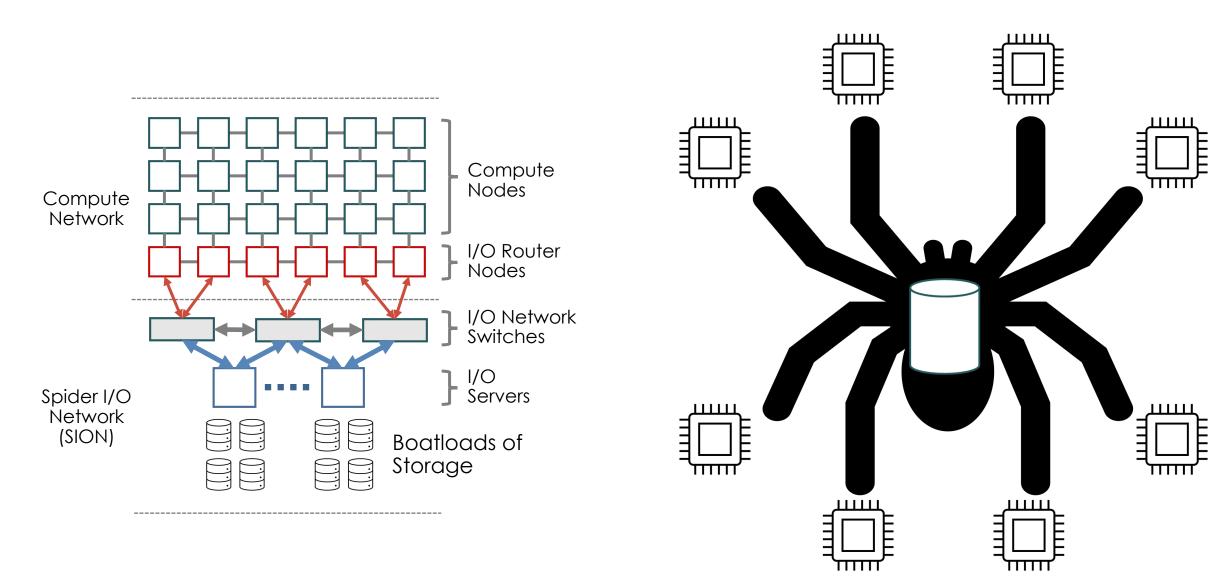


Why Center-wide Shared File Systems?

- Historically, supercomputers were deployed with a tightlycoupled HPC storage solution
- This approach has several drawbacks:
 - 1. Storage is expensive \$\$\$\$ (can be up to a 1/3 of HPC system cost)
 - 2. Many storage systems == more administrator work & user confusion
 - 3. Increases large-scale data movement
 - e.g., to move simulation results to a data analysis cluster
 - 4. Tight-coupling often meant system downtimes made storage unavailable



Spider - An Architecture for a Center-wide Shared FS



CAK RIDGE National Laboratory

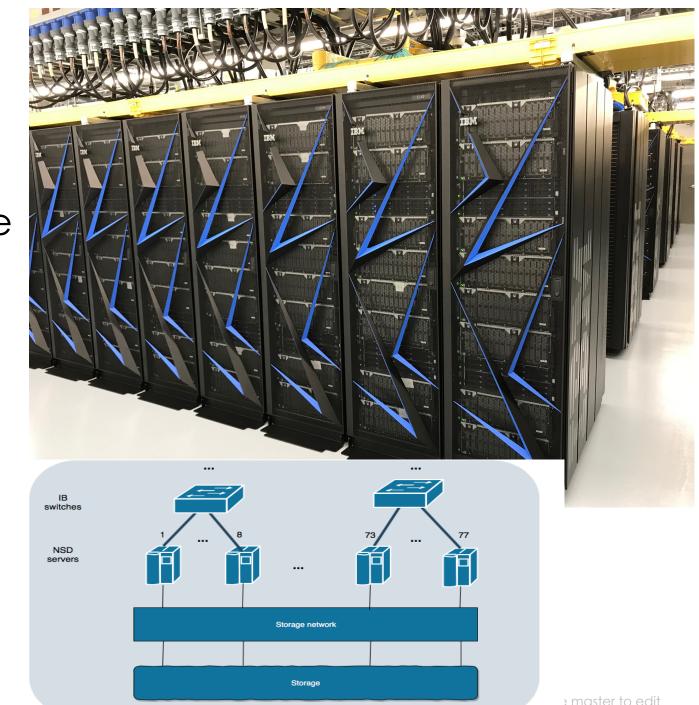
Spider Through the Years

	FS Type	Leadership System	# of Clients (est.)	Capacity	# Disks	Hero Bandwidth (measured)
Spider 1 (2008)	Lustre	Jaguar/Titan	26,000	10 PB	ŚŚ	240 GB/s
Spider 2 (2013)	Lustre	Titan	26,000	32 PB	~20K	1.2 TB/s
Spider 3 (2018)	Spectrum Scale	Summit/Frontier	12,000	250 PB	~30K	2.5 TB/s
Spider 4 (2023)	Lustre	Frontier	10,000	700 PB	~53K	4.7 TB/s (Capacity Tier)



Spider3 – Alpine (EOL)

- 250 PB usable capacity
- 2.5 TB/s sequential read/write
- 2.2 TB/s random read/write



- 77 IBM Elastic Storage Server (ESS) GL4s
- 32,494 10TB NL-SAS drives

COAK RIDGE National Laboratory

Spider4 - Orion

- 679 PB usable capacity
- 40 metadata servers
- 450 storage servers
- 160 Router nodes



• 5 rows of racks _

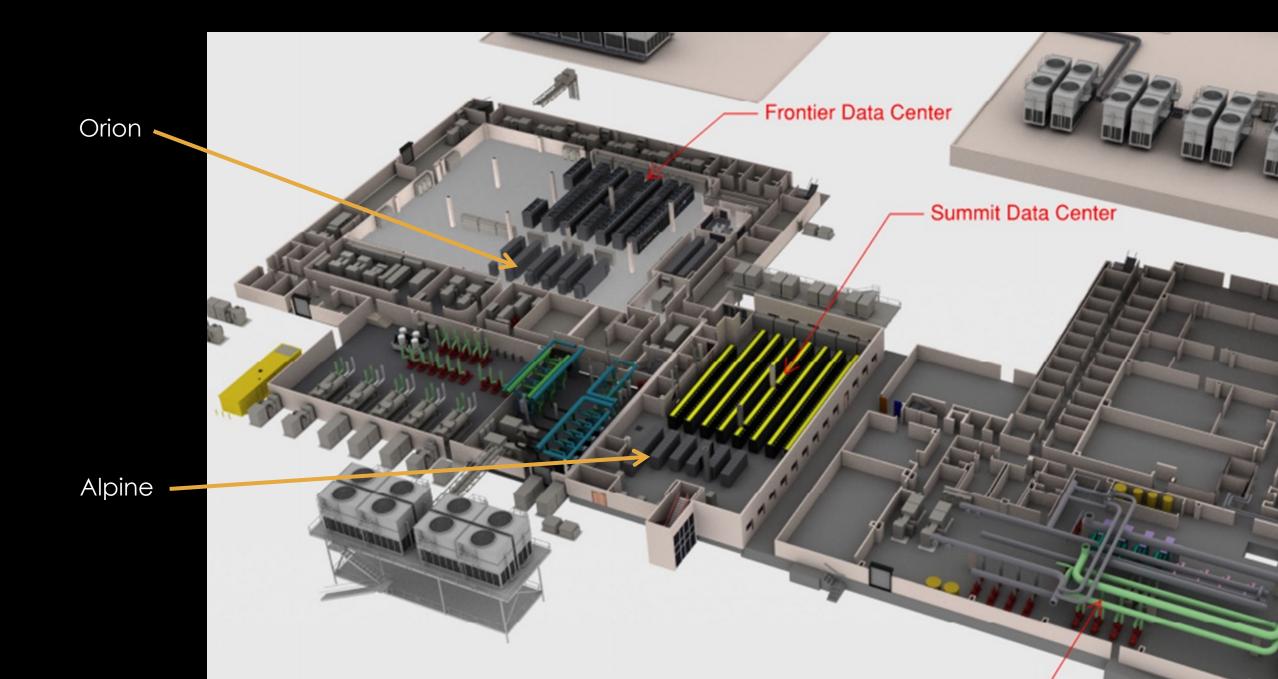
i Ori Oris

• +1 mgmt rack

COAK RIDGE National Laboratory

Orion-etc-core2 Orion-eth-core1	Orion-oss11d[1-2]	Orion-oss12d[1-2]	Orion-oss13d[1-2]	Orion-oss14d[1-2]	Orion-oss15d[1-2]	Orion-eth1f Orion-eth1e	Orion-oss16d[1-2]	Orion-oss17d[1-2]	Orion-oss18d[1-2]	Orion-oss19d[1-2]
Orion-mgmt[1-2]	Orion-oss11c[1-2]	Orion-oss12c[1-2]	Orion-oss13c[1-2]	Orion-oss14c[1-2]	Orion-oss15c[1-2]	Orion-eth1d Orion-eth1c	Orion-oss16c[1-2]	Orion-oss17c[1-2]	Orion-oss18c[1-2]	Orion-oss19c[1-2]
1U Free Orion-Ids-core2	Orion-oss11b[1-2]	Orion-oss12b[1-2]	Orion-oss13b[1-2]	Orion-oss14b[1-2]	Orion-oss15b[1-2]	Orion-eth1b Orion-eth1a	Orion-oss16b[1-2]	Orion-oss17b[1-2]	Orion-oss18b[1-2]	Orion-oss19b[1-2]
Orion-tds-core1 Orion-CDS-mgmt2	Orion-oss11a[1-2]	Orion-oss12a[1-2]	Orion-oss13a[1-2]	Orion-oss14a[1-2]	Orion-oss15a[1-2]	Orion-cds-eth1b Orion-cds-eth1a	Orion-oss16a[1-2]	Orion-oss17a[1-2]	Orion-oss18a[1-2]	Orion-oss19a[1-2]
Orion-CDS-mgmt1 FM02 FM01 Orion-mgmt-eth2	Orion-enc-11a2	Orion-enc-12a2	Orion-enc-13a2	Orion-enc-14a2	Orion-enc-15a2	Orion-ss-sw1p Orion-ss-sw1o Orion-ss-sw1n Orion-ss-sw1m	Orion-enc-16a2	Orion-enc-17a2	Orion-enc-18a2	Orion-enc-19a2
Orion-mgmt-eth1 3U Free	Orion-enc-11a1	Orion-enc-12a1	Orion-enc-13a1	Orion-enc-14a1	1 Orion-enc-15a1	Orion-ss-sw11 Orion-ss-sw1k Orion-ss-sw1j Orion-ss-sw1j	Orion-enc-16a1	Orion-enc-17a1	Orion-enc-18a1	Orion-enc-19a1
Orion-CDS3 Orion-CDS2 Orion-CDS1 Orion-util12	Orion-enc-11b2	Orion-enc-12b2	Orion-enc-13b2	Orion-enc-14b2	Orion-enc-15b2	Orion-ss-sw1h Orion-ss-sw1g Orion-ss-sw1f Orion-ss-sw1f	Orion-enc-16b2	Orion-enc-17b2	Orion-enc-18b2	Orion-enc-19b2
Orion-util11 Orion-util10 Orion-util9 Orion-util8	Orion-enc-11b1	Orion-enc-12b1	Orion-enc-13b1	Orion-enc-14b1	Orion-enc-15b1	Orion-ss-sw1d Orion-ss-sw1c Orion-ss-sw1b Orion-ss-sw1b	Orion-enc-16b1	Orion-enc-17b1	Orion-enc-18b1	Orion-enc-19b1
Orion-util7 Orion-util6 Orion-util5 Orion-util4	Orion-enc-11c2	Orion-enc-12c2	Orion-enc-13c2	Orion-enc-14c2	Orion-enc-15c2	Orion-rtr1h[1-4] Orion-rtr1g[1-4]	Orlon-enc-16c2	Orion-enc-17c2	Orion-enc-18c2	Orion-enc-19c2
Orion-util3 Orion-util2 Orion-util1 1U Free	Orion-enc-11c1	Orion-enc-12c1	Orion-enc-13c1	Orion-enc-14c1	Orion-enc-15c1	Orion-rtr1f(1-4)	Orion-enc-16c1	1 Orion-enc-17c1	Orion-enc-18c1	Orion-enc-19c1
rion-tds-mgmt[1-2]	1 Orion-enc-11d2 1	Orion-enc-12d2	1 Orion-enc-13d2	1 Orion-enc-14d2	Orion-enc-15d2	Orion-ttr1d[1-4]	Orion-enc-16d2	1 Orion-enc-17d2	1 Orion-enc-18d2	Orion-enc-19d2
Drion-tds-mds[1-2]	Orion-enc-11d1	Orion-enc-12d1	Orion-enc-13d1	Orion-enc-14d1	Orion-enc-15d1	Orion-rtr1b[1-4]	Orion-enc-16d1	Orion-enc-17d1	Orion-enc-18d1	Orion-enc-19d1
Orion-tds-enc-1a2	Orion-enc-11e2	Orion-enc-12e2	Orion-enc-13e2	Orion-enc-14e2	Orion-enc-15e2	Orion-mds1d[1-2]	Orion-enc-16e2	Orion-enc-17e2	Orion-enc-18e2	Orion-enc-19e2
Orion-tds-enc-1a1	Orion-enc-11e1	Orion-enc-12e1	Orion-enc-13e1	Orion-enc-14e1	Orion-enc-15e1	Orion-mds1b[1-2]	Orion-enc-16e1	Orion-enc-17e1	Orion-enc-18e1	Orion-enc-19e1
53 Orion MGMT / TDS	CB25 Orion-11	CA25 Orion-12	BZ25 Orion-13	BY25 Orion-14	BX25 Orion-15	BW25 Orion1-Infra	BV25 Orion-16	BU25 Orion-17	BT25 Orion-18	BS25 Orion-19

OLCF Computational Facilities





Lustre and Orion

- Architecture and Features of Lustre
- OLCF Orion Details

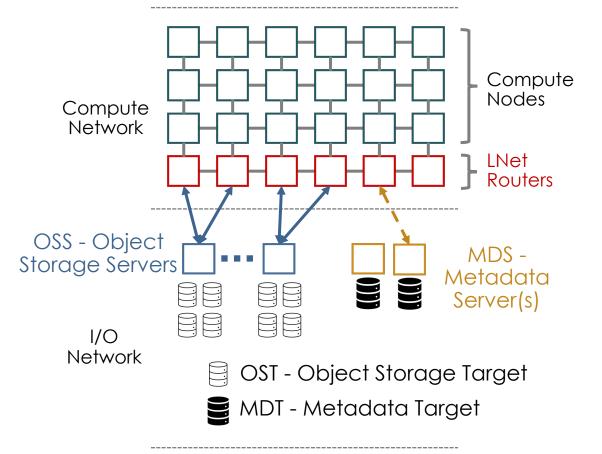


Lustre File System Architecture

- Two Types of Servers
 - Metadata Servers (MDS): maintain file system hierarchy, serve file properties/metadata
 - Object-Storage Servers (OSS): serve file data
- Clients

CAK RIDGE National Laboratory

- talk to MDS to navigate FS, retrieve stats, and locate file extents
- talk directly to OSS to read/write file extents

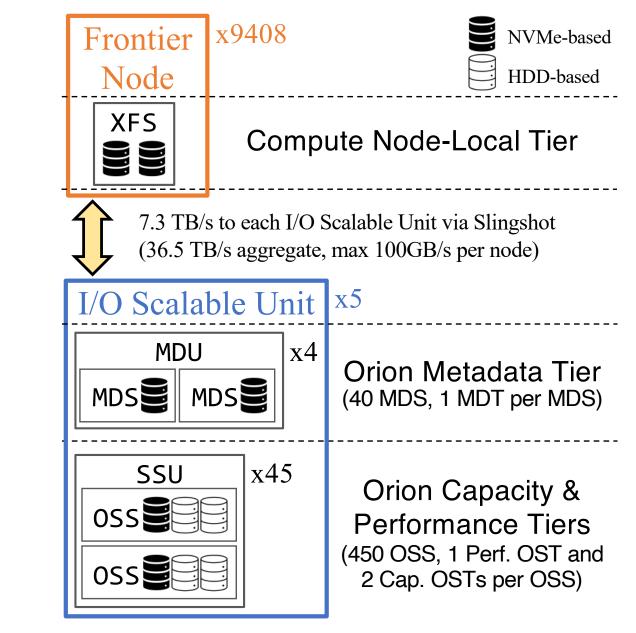


Orion Tiered Architecture

- Capacity Tier:
 - 679 PB
 - RD/WR: 5.5/4.6 TB/s
 - 47,700 18 TB HDD
- Performance Tier:
 - 11.5 PB
 - RD/WR: 10 TB/s
 - 5,400 3.2 TB NVMe
- Metadata Tier:
 - 10 PB

COAK RIDGE National Laboratory

- RD/WR: 0.8/0.4 TB/s
- 480 30.7 TB NVMe



Lustre File Data Management

- Terminology
 - Lustre Stripe Size (SS): data object size used to spread data round-robin across OSTs
 - Lustre Stripe Count (SC): number of OSTs used for striping a given file
- Example: 1 GiB file, SS=1 MiB, SC=8
 - File divided into 1024 data objects, 128 objects assigned to each OST
- Both SS and SC are user-controllable in Lustre

 either at a directory level, or per-file
- Facilities do their best to set reasonable defaults, but use cases very dramatically, and can lead to decreased performance



New Features of Lustre

- Data on MDT (DoM)
 - for very small files, store the file data on MDT with its metadata
- Distributed Namespace Extension (DNE)
 - ability to employ more than one MDT to manage directories in a single file system (DNE1), or to stripe directory entries across MDTs (DNE2)
- Progressive File Layout (PFL)
 - a composite layout that uses different stripe sizes (and possibly widths) for predefined regions of a file
 - e.g., 16 KiB for first [0, 1MiB), 1 MiB for [1 MiB, 1 GiB), 64 MiB for [1 GiB, EOF)
- Self-Extending Layout

COAK RIDGE National Laboratory

extension to PFL that avoids OST out-of-space conditions for small stripe sizes

Increasing User Satisfaction on Orion

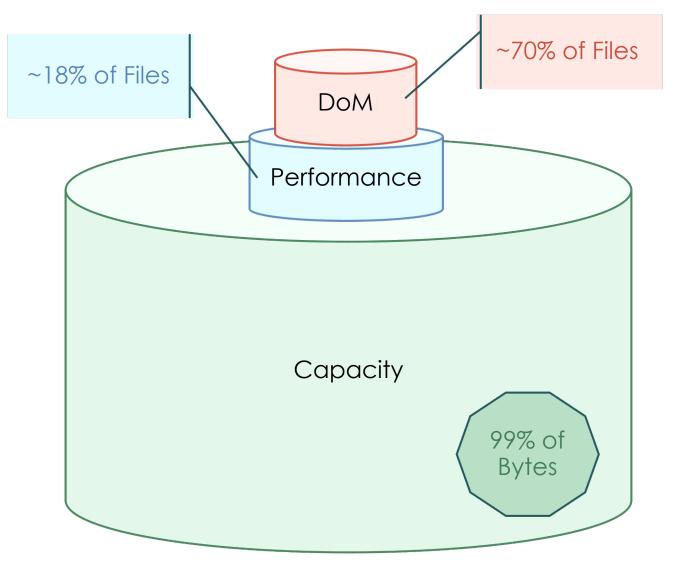
- Despite the somewhat complex tiered architecture of Orion, the design goals are:
 - 1. ease-of-use
 - 2. good I/O performance for common workloads
- OLCF has a large corpus of I/O profiling data collected on Summit/Alpine using Darshan
 - Analysis of this profile data suggests that a single default progressive file layout can achieve both goals



Default Progressive File Layout on Orion

- First 256 KiB of each file on Metadata Tier
 - SS=256KiB, SC=1
- Next 8 MiB of each file on Performance Tier
 - SS=1MiB, SC=1
- Next 128 GiB of each file on Capacity Tier
 - SS=1MiB, SC=1
- Rest on Capacity Tier
 - SS=1MiB, SC=8

COAK RIDGE National Laboratory



DNE for Metadata Isolation on Orion

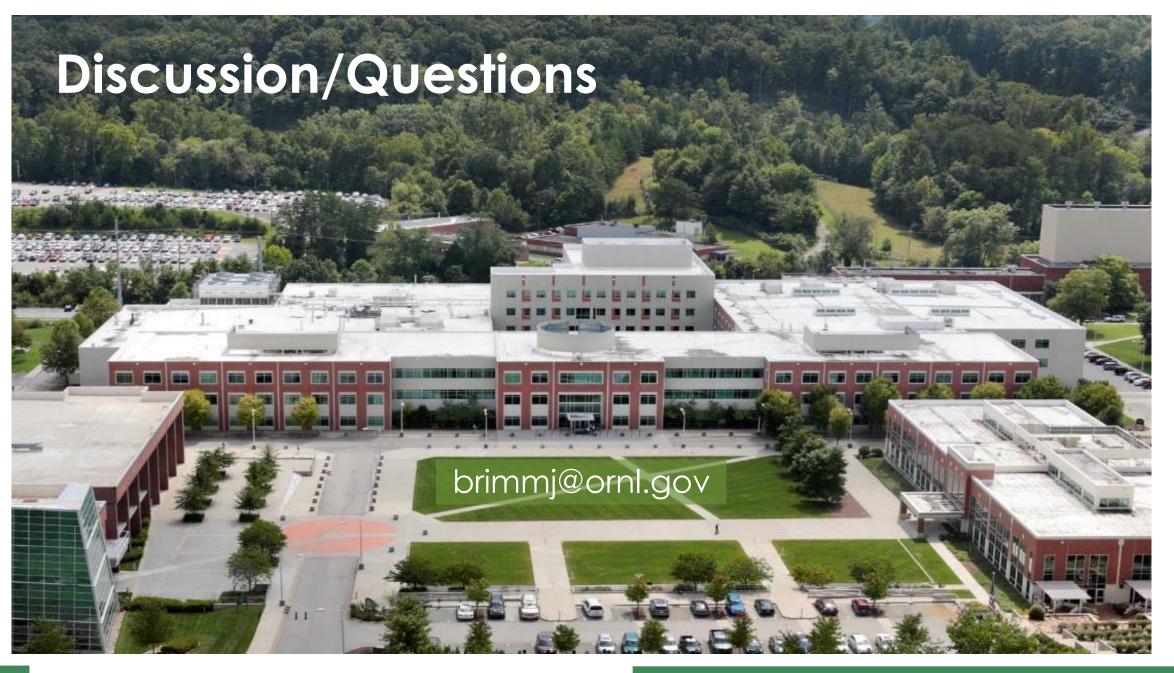
- Each allocation/project is assigned to exactly one of the 40 metadata servers
- Potential Benefits
 - Increased metadata caching performance
 - Decreased server contention from concurrent metadata access workloads from other projects
 - Easier to load-balance metadata-intensive projects



Pre-Production Application I/O Performance on Orion

- Default progressive file layout is good for file-per-process
 - WarpX (ADIOS) File-per-process @ 4,096 nodes (32k processes, 1.5 GiB/process) achieved ~7.9 TiB/sec for simulation output writing
 - GTC (ADIOS) File-per-process @ 2,048 nodes (16k processes, 2 GiB/process) achieved ~5.2 TiB/sec for checkpointing three datasets
- But not so great for large single-shared-file
 - Flash-X (HDF5) Shared-file @ 512 nodes (28k processes, 29 GiB/node) got only 6 GiB/sec
 - Using Capacity tier only with wide-striping improved this by 20x





COAK RIDGE

24

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