OAK RIDGE National Laboratory

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

Overview and Goals for this Lecture

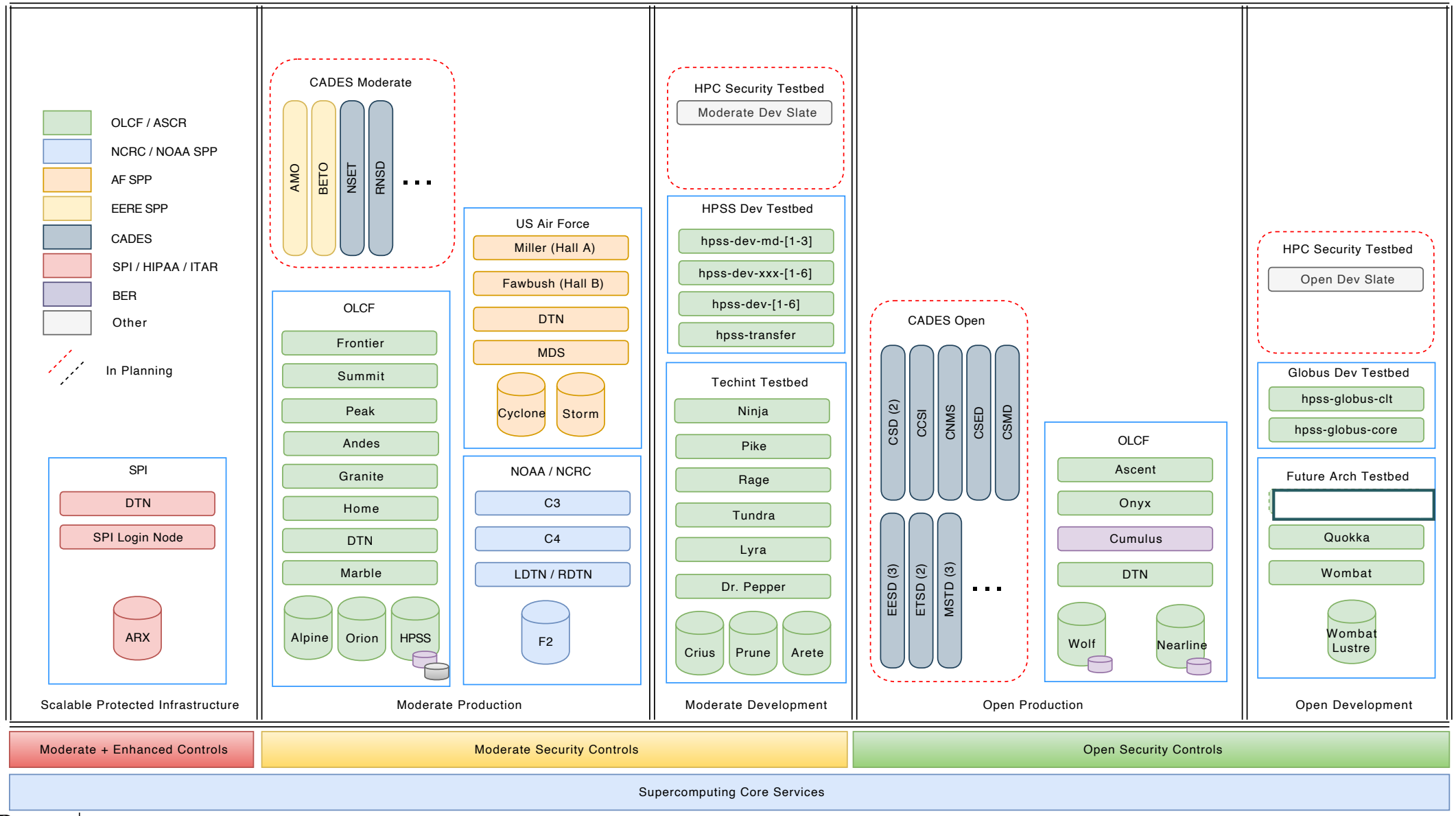
1. Quick background on NCCS and OLCF approach to HPC storage
2. High-level overview of OLCF's center-wide shared storage system
3. Deep dive on Lustre and Orion

OAK RIDGE National Laboratory

HPC Storage @ NCCS and OLCF

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

National Center for Computational Sciences



OAK RIDGE National Laboratory | LEADERSHIP COMPUTING FACILITY

Last updated: Ryan Adamson 2021-09-10

NCCS HPC Storage Strategy

		OLCF	BER	AirForce	NOAA - NCRC
Faster ->	**Job-term (24 hours or less)**	Summit Frontier			
	Short-term (less than 90 days)	Alpine/Orion Arx Wolf	Wolf	Storm Cyclone	F2
	Medium-Term (90 days to 1-3 years)	HPSS/Themis		N/A	N/A
<- Slower	**Long-term (90 days to 20 years)**			N/A	N/A
	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

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

F2
- NCRC/Lustre-2.12
- 40 PB
- 45 GB/s r/w

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

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

Production Archive Filesystems

HPSS

- HPSS-7.2
- 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

OAK RIDGE National Laboratory

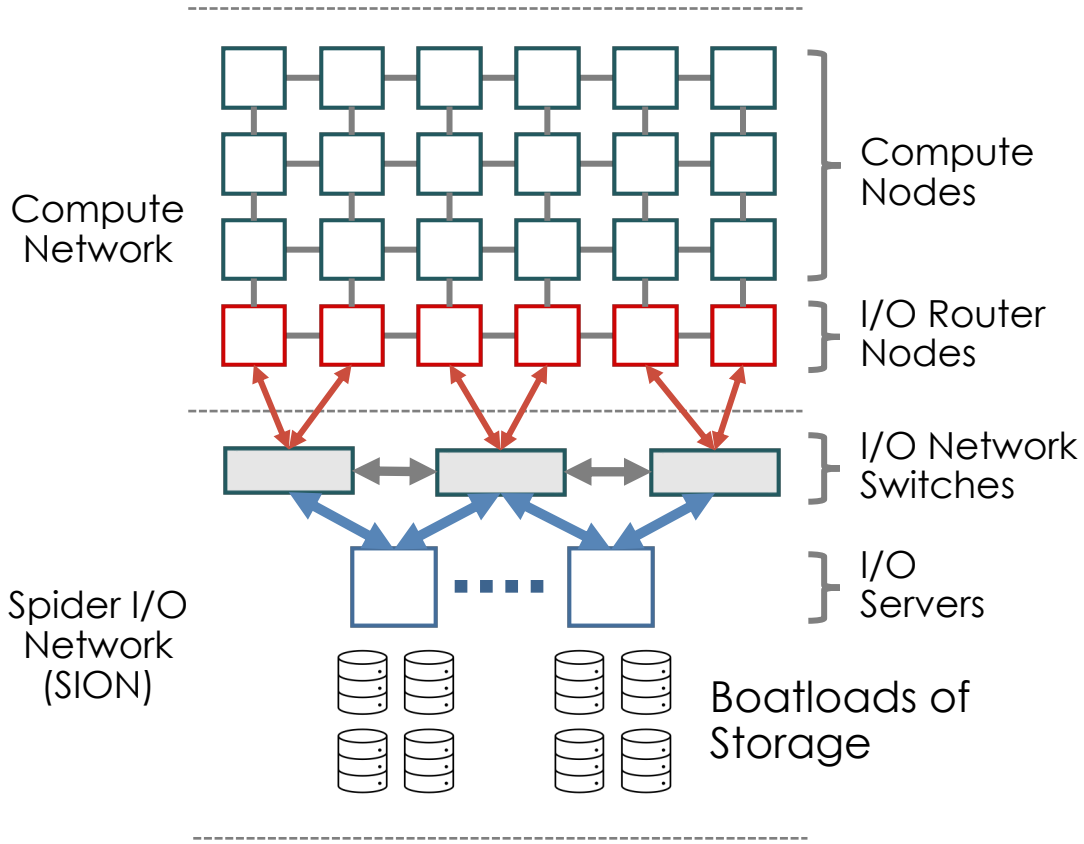
The OLCF Center-wide Shared File System

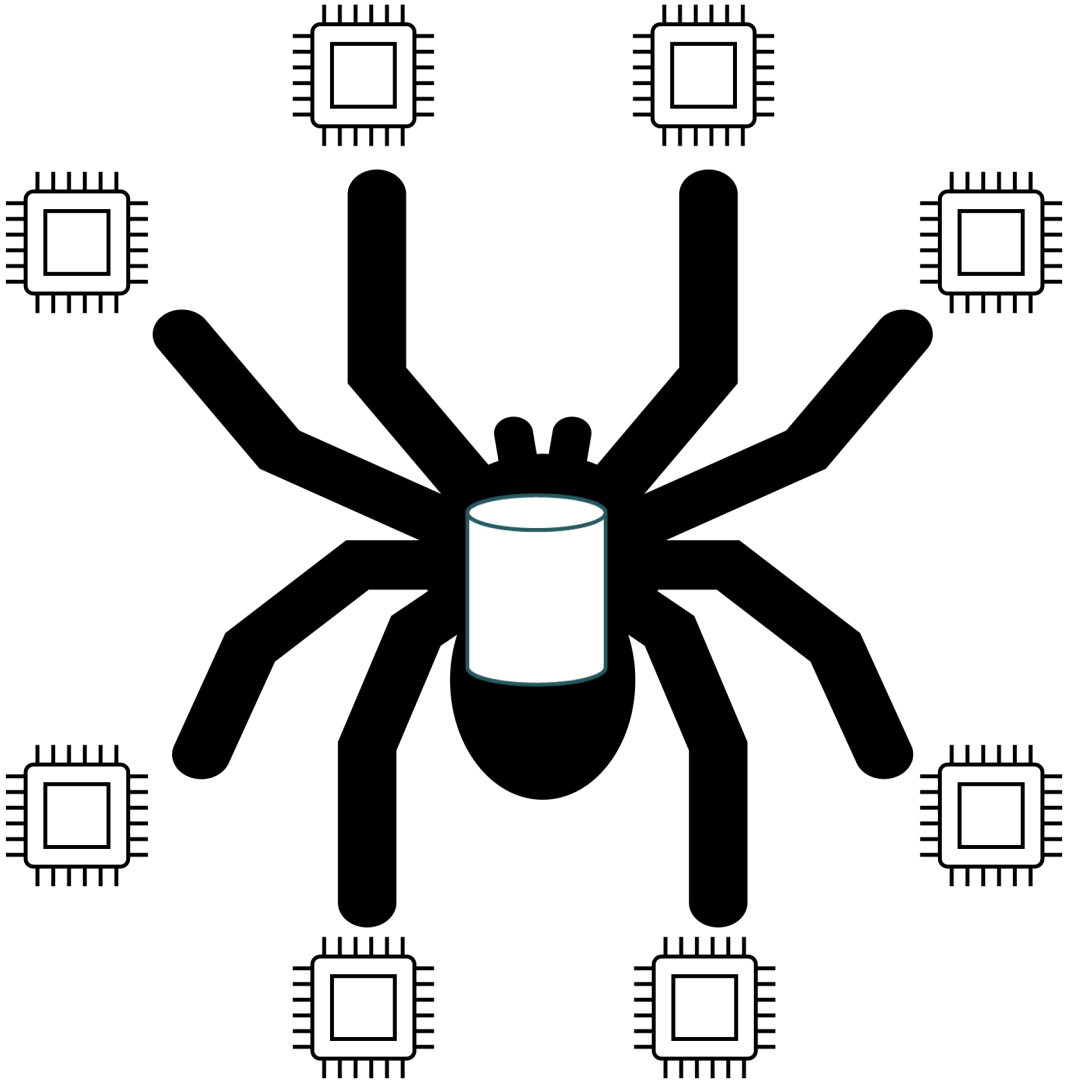
- Why center-wide?
- Spider Architecture and History
- Production File Systems

Why Center-wide Shared File Systems?

- Historically, supercomputers were deployed with a tightly-coupled 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





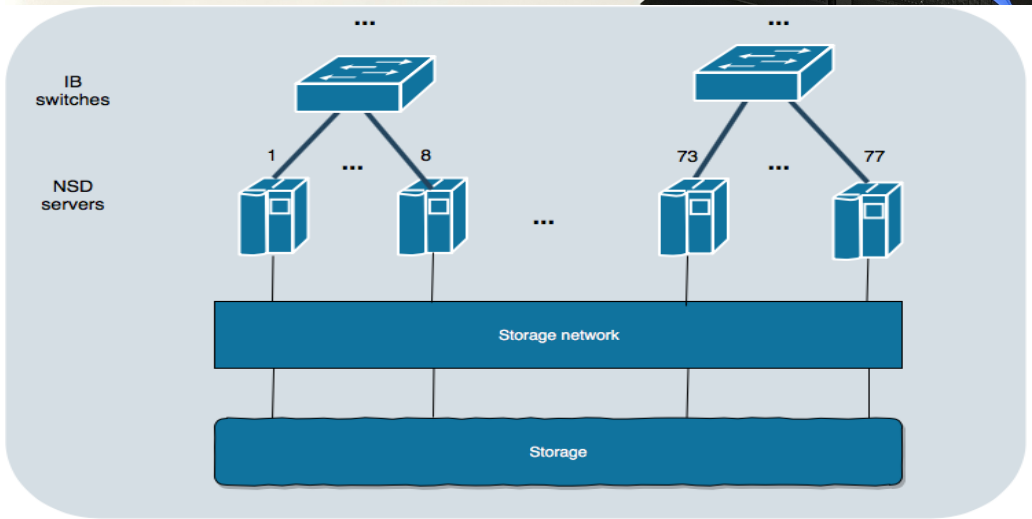
Spider Through the Years

	FS Type	Leadership System	# of Clients (est.)	Capacity	# Disks	Hero Bandwidth (measured)
Spider 1 (2008)	Lustre	Jaguar/Titan	26,000	10 PB	??	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

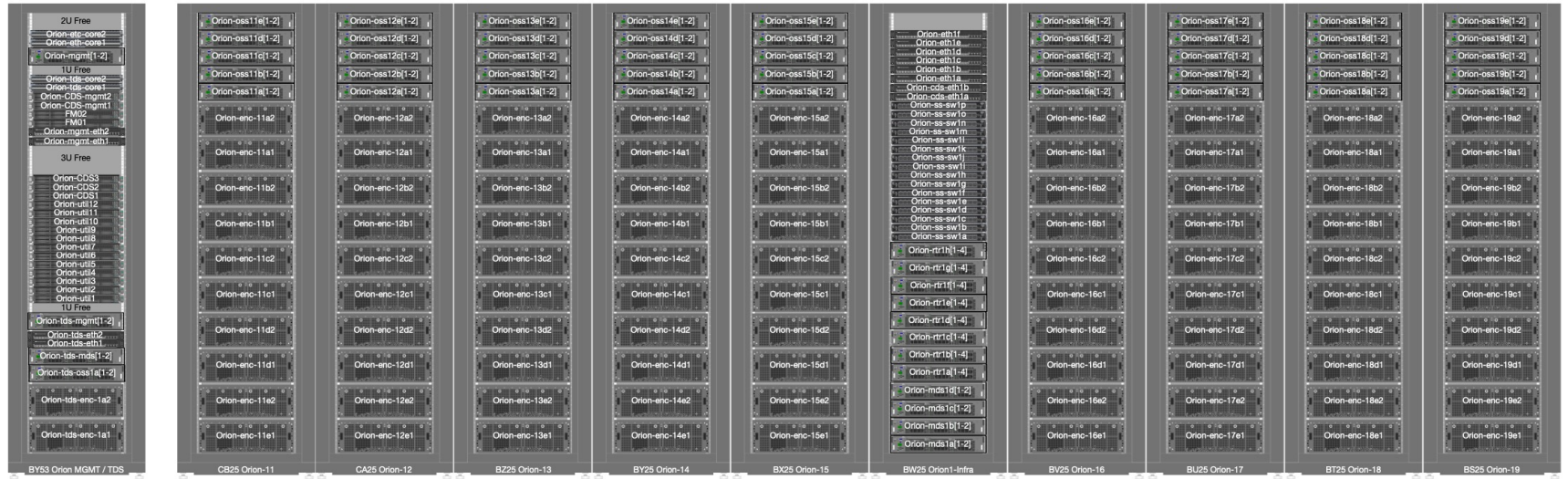
- 77 IBM Elastic Storage Server (ESS) GL4s
- 32,494 10TB NL-SAS drives



Spider4 - Orion

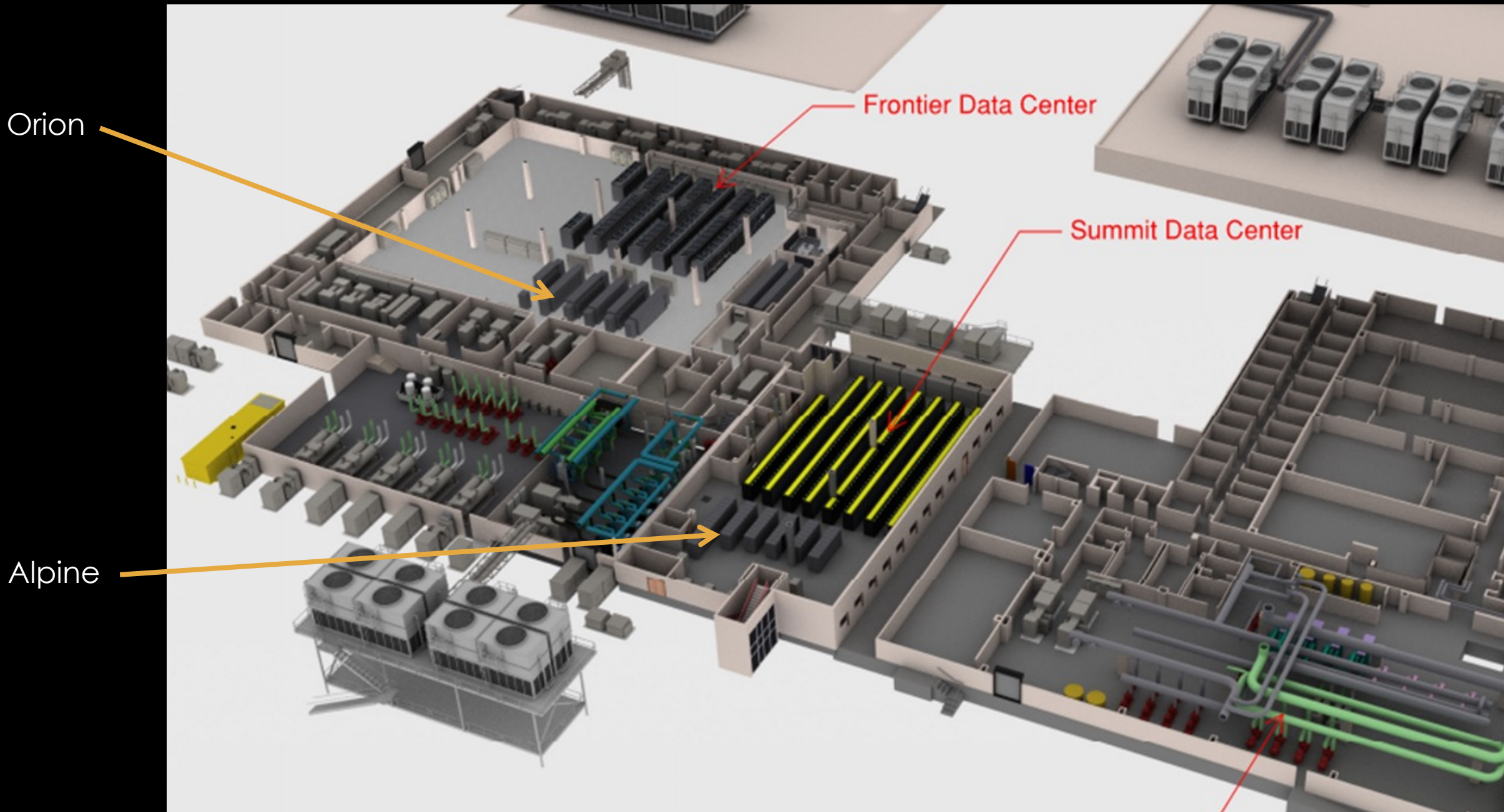
- 679 PB usable capacity
- 40 metadata servers
- 450 storage servers
- 160 Router nodes

- 5 rows of racks
- +1 mgmt rack



OAK RIDGE National Laboratory | LEADERSHIP COMPUTING FACILITY

OLCF Computational Facilities



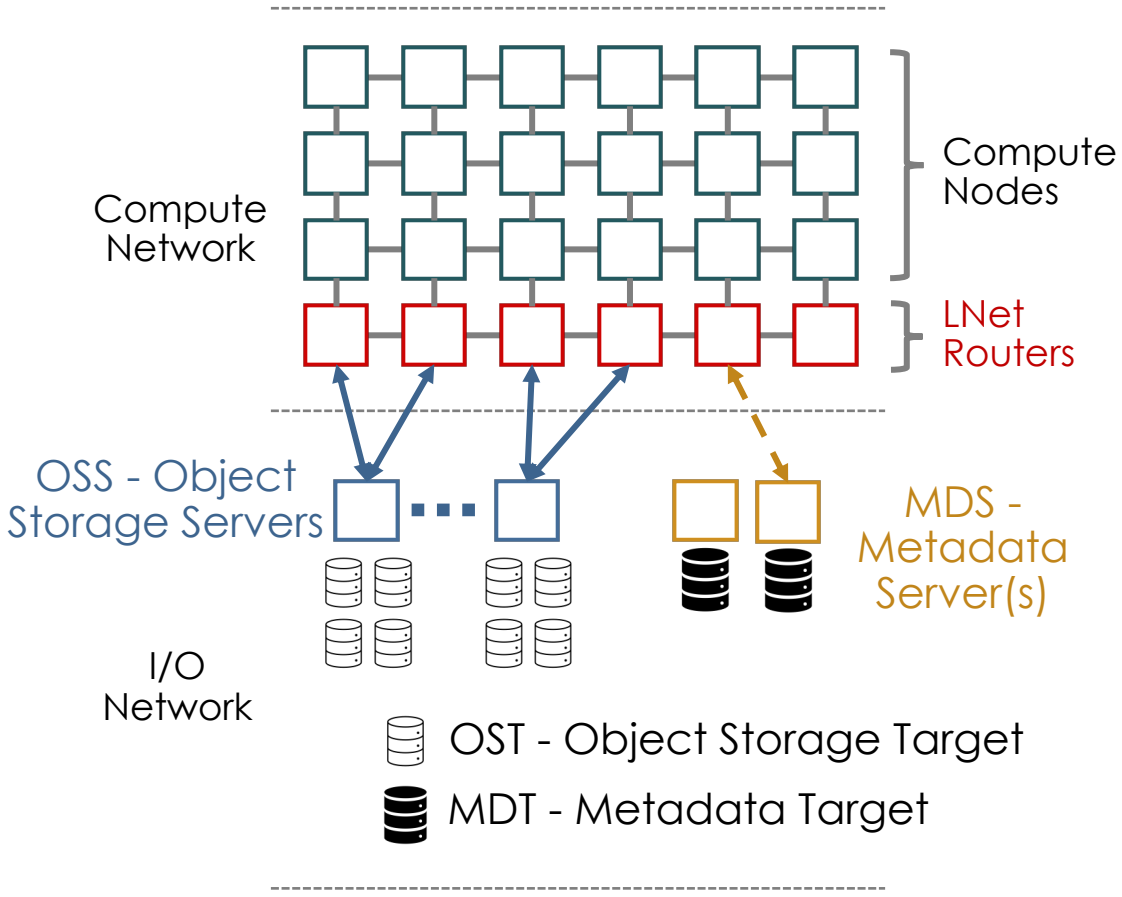
OAK RIDGE National Laboratory

Lustre and Orion

- Architecture and Features of Lustre
- OLCF Orion Details

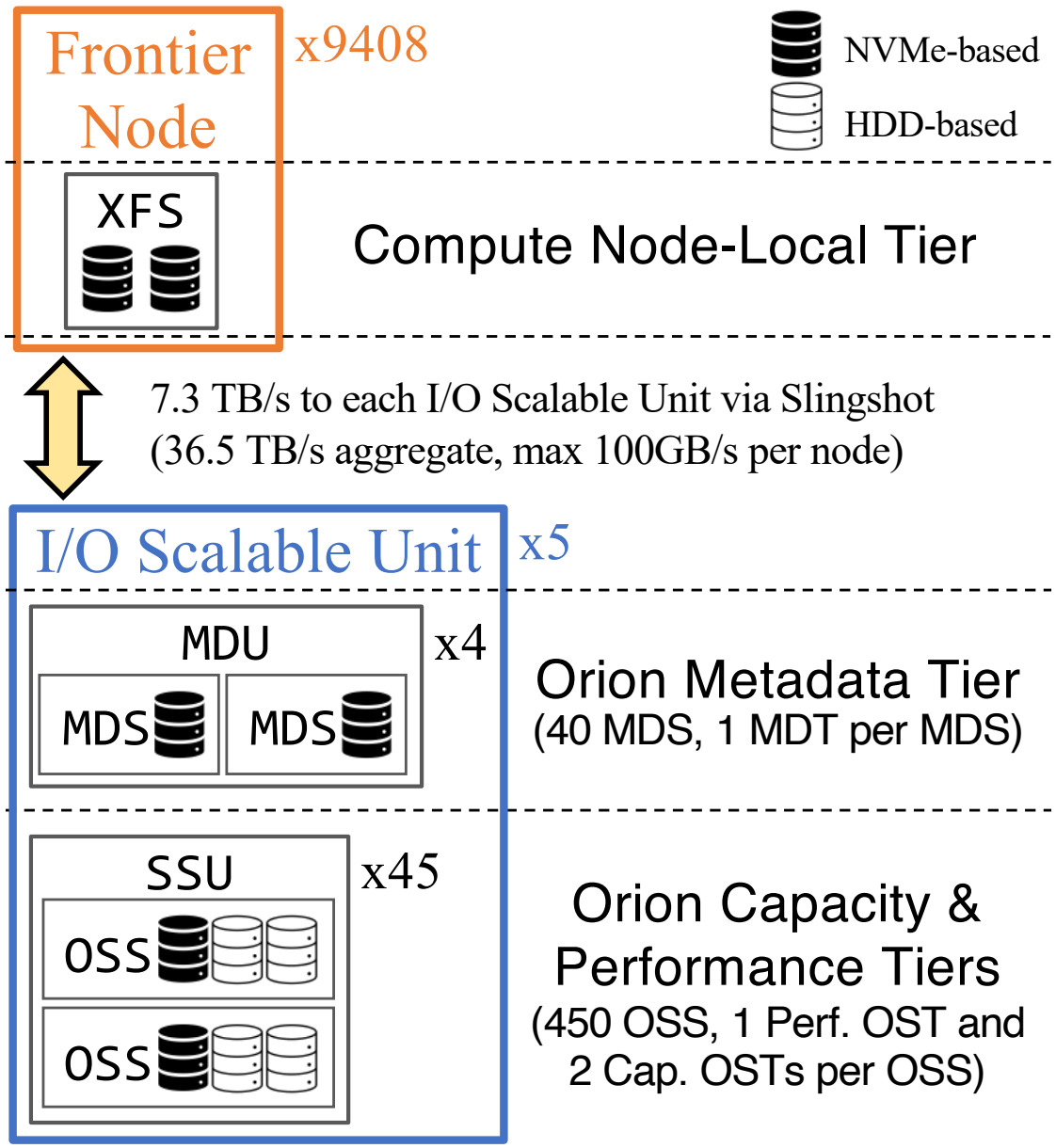
Lustre File System Architecture

- Two Types of Servers
  - Metadata Servers (MDS): maintain file system hierarchy, serve file properties/metadata
  - Object-Storage Servers (OSS): serve file data
- Clients
  - 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
  - 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
  - extension to PFL that avoids OST out-of-space conditions for small stripe sizes

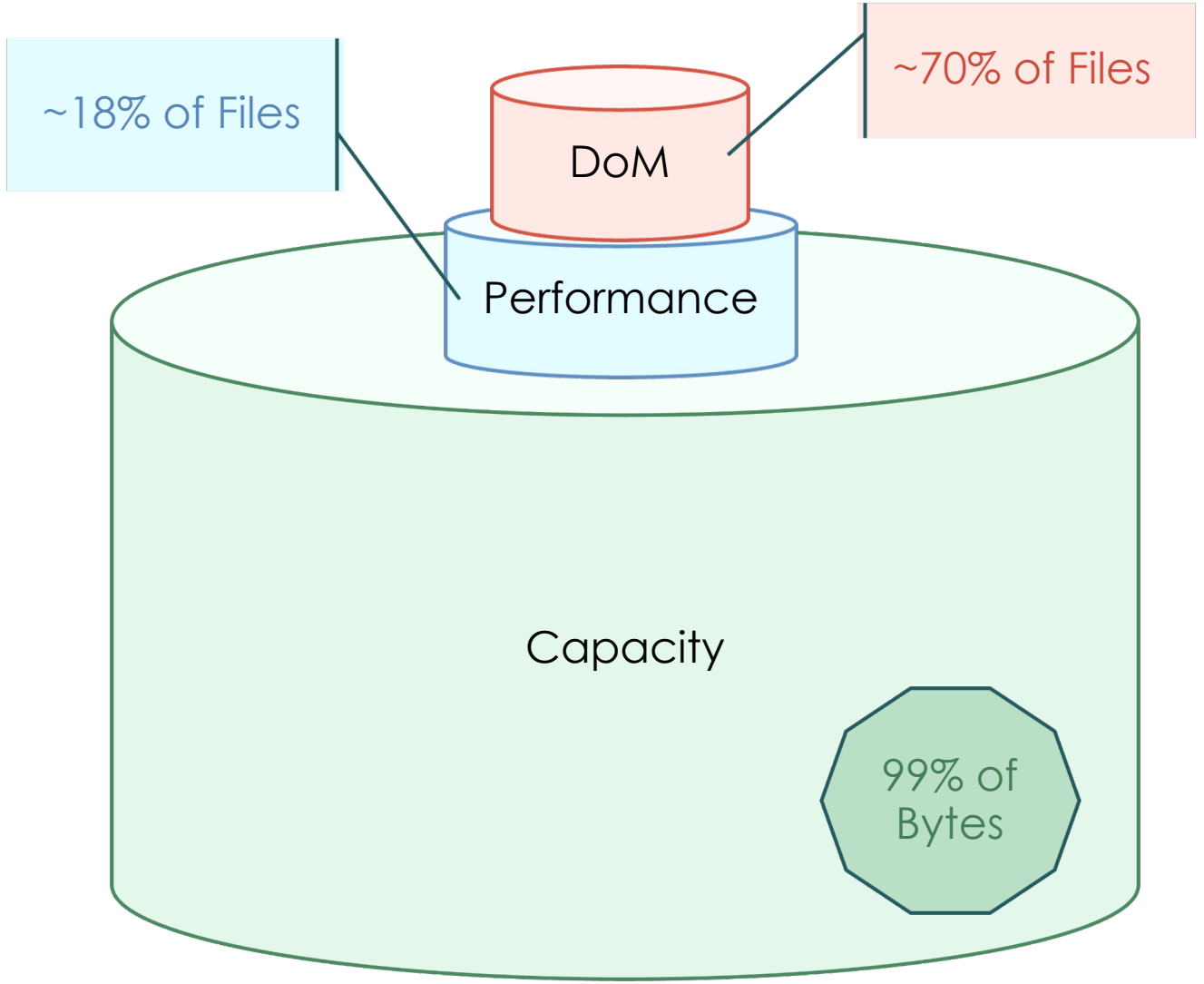
Increasing User Satisfaction on Orion

- 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



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

Discussion/Questions

brimmj@ornl.gov

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