OAK RIDGE National Laboratory

Deploying a Parallel File System for the World's First Exascale Supercomputer

Jesse Hanley

Dustin Leverman

CUG'2023

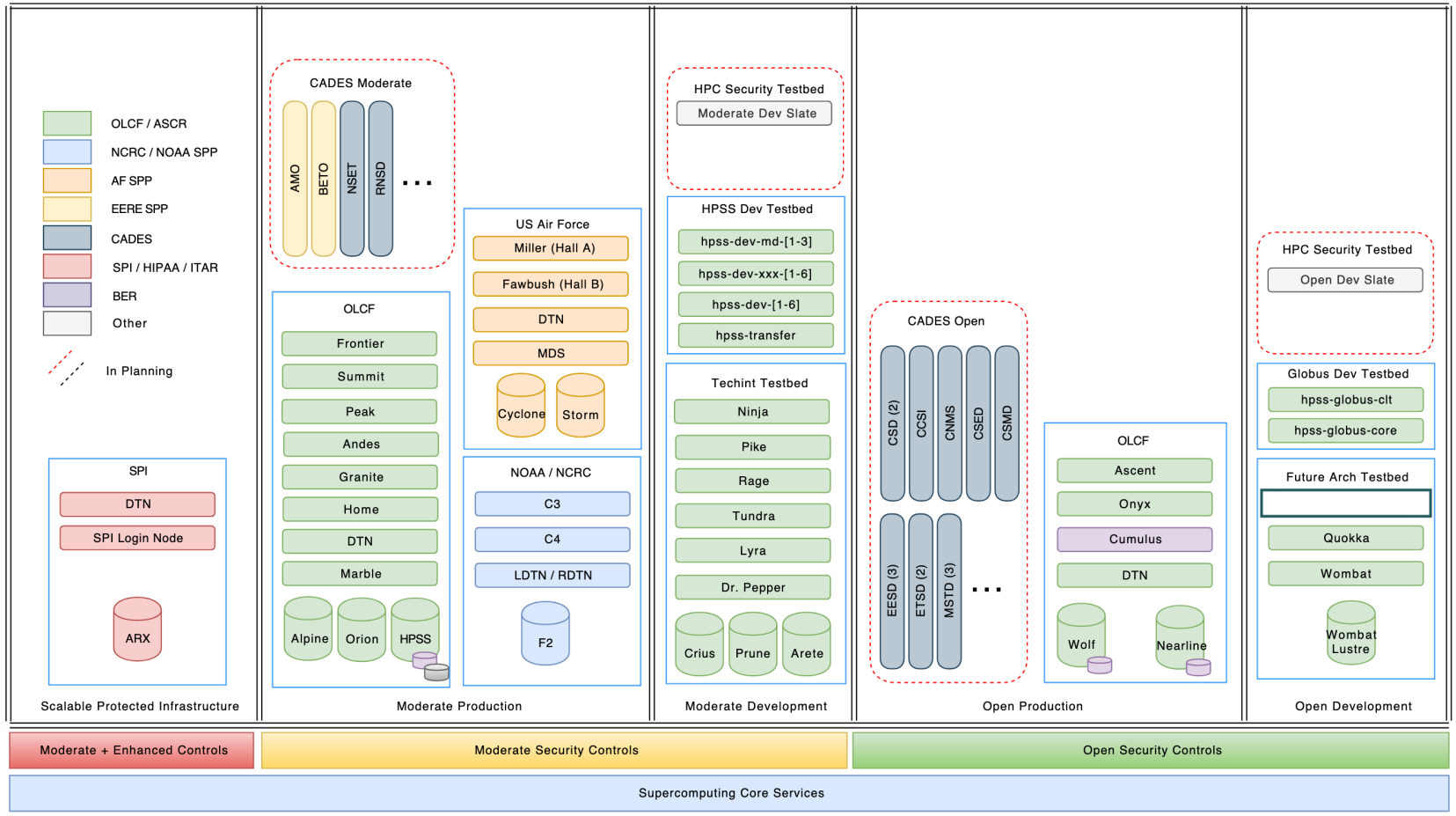
ORNL is managed by UT-Battelle LLC for the US Department of Energy

U.S. DEPARTMENT OF ENERGY

Introduction

- Orion Overview
- System Configuration
- Overview/Subset of Acceptance Process

NCCS Systems



Orion Overview

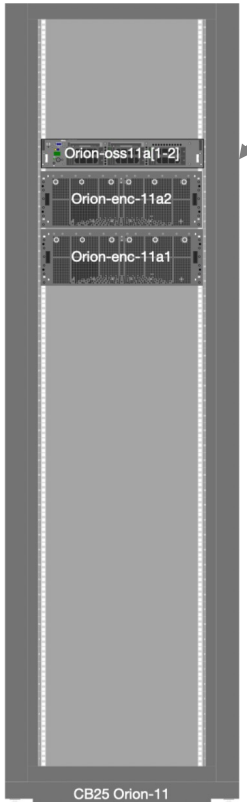
2x Management Nodes	2x MGS	40x MDS	450x OSS
160x LNET Router Nodes	12x Utility Nodes	80x Slingshot Switches	35x Ethernet Switches

Nodes are grouped into Storage Scalable Units (SSU)

SSUs are grouped (with networking) into Storage Scalable Clusters (SSC)

Orion Building Blocks

SSU

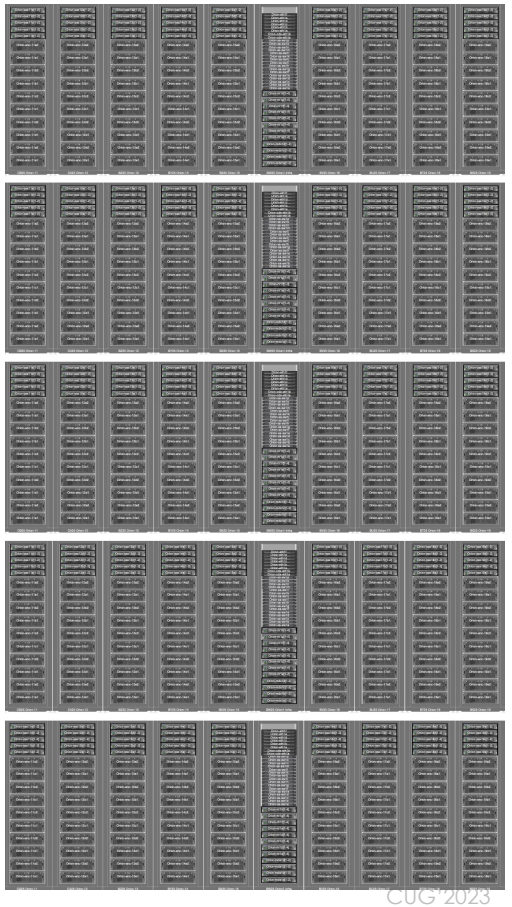


2x OSS nodes
- 24x 3.84TB NVMe
- Each w/ 256GB DDR4
- 32-core AMD EPYC gen2
- 2x Brazos SS ports

2x 106-bay SAS enclosures
- 212x 18TB PMR HDD

5x SSC

Full System

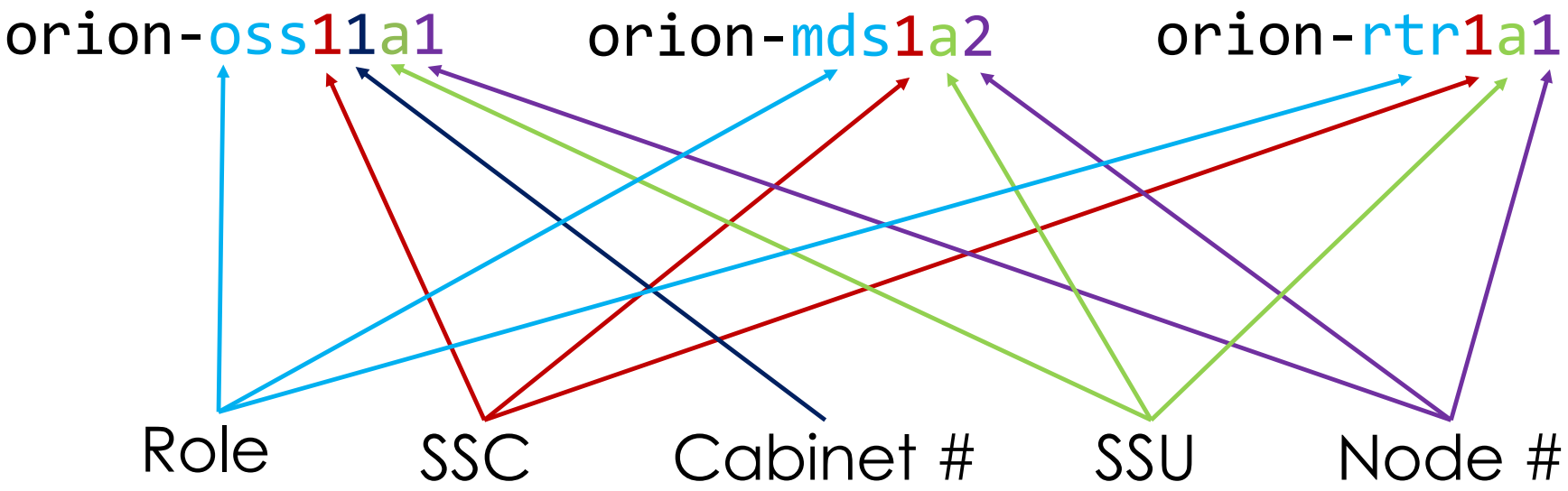


45x SSUs
32x LNET router nodes
8x MDS nodes
- Identical to OSS except for NVMe are 30TB

5x Ethernet Management switches
1x SS "group" of 16x switches

SSC

Orion Host Naming Convention

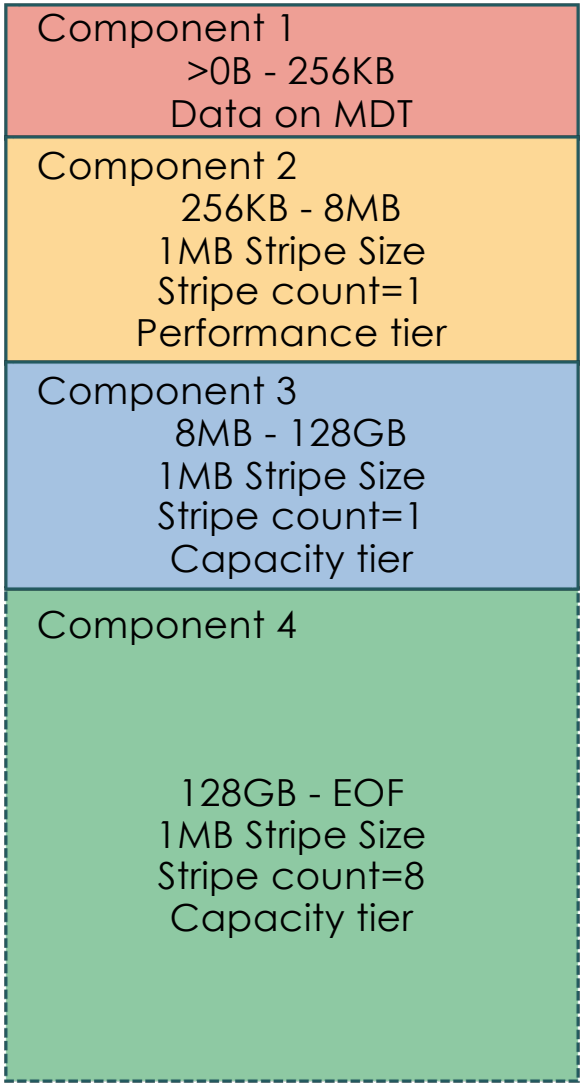


Tiering

- 9.7 PB NVMe-based MDT storage
  480 drives
- 11.4 PB NVMe-based OST storage
  5,400 drives
- 667.6 PB HDD-based OST storage
  47,700 drives

- Uses Lustre Progressive File Layout (PFL), Data on MDT (DoM), and Self-Extending Layouts (SEL);
- Utilizing DNEp1 and "randomly" assigning projects across 40 MDTs

```
/usr/bin/lfs setstripe
-E 256K -L mdt
-E 8M -c 1 -S 1M -p performance -z 64M
-E 128G -c 1 -S 1M -z 16G -p capacity
-E -1 -z 256G -c 8 -S 1M -p capacity /lustre/orion/
```



Management Stack

- Cluster Provisioning using Anchor
  - `dnsmasq` updated dynamically based on switch-port configs
  - `matchbox` acts as node classifier
  - `squashfs` image distributed to nodes (compressed; read-only)
  - mounted with Dracut module using read-write overlay
- ClusterStor recipe for Data Path
  - Base RHEL image
  - HPE provided kernel, Lustre, ZFS, firmware, HA, etc…
- Redundant (hot/cold) management servers
  - System boot time ~7min

System Monitoring

Examples (non-exhaustive)

Hardware

- IPMI
- SAS Health
- HDD Enclosures
- NVMe
- Disk
- Firmware Versioning

Software

- LNET
- Normal Linux daemons
  - `NTP`
  - `crond`
  - `syslog`
  - …
- Configuration management run history

HPE Tooling

- Disk Monitoring
- Disk Watch Daemon
- High Availability
- Slingshot

Namespace health

- MDT, Perf, and Cap tier utilization
- `ls` timer
- OST states
  - D - degraded
  - N – no-precreate
  - R – read-only
  - I – out of space
  - S – out of inodes

Goal of monitoring: monitor and alert appropriately to detect issues before users do

- Involves alerting differently depending on if during business hours or after hours

Acceptance Overview

Phase 1: System install and checkout
Phase 2: Single-unit testing
Phase 3: Scale up
Phase 4: Full system testing

Multiple acceptance phases, each phase can include the following:

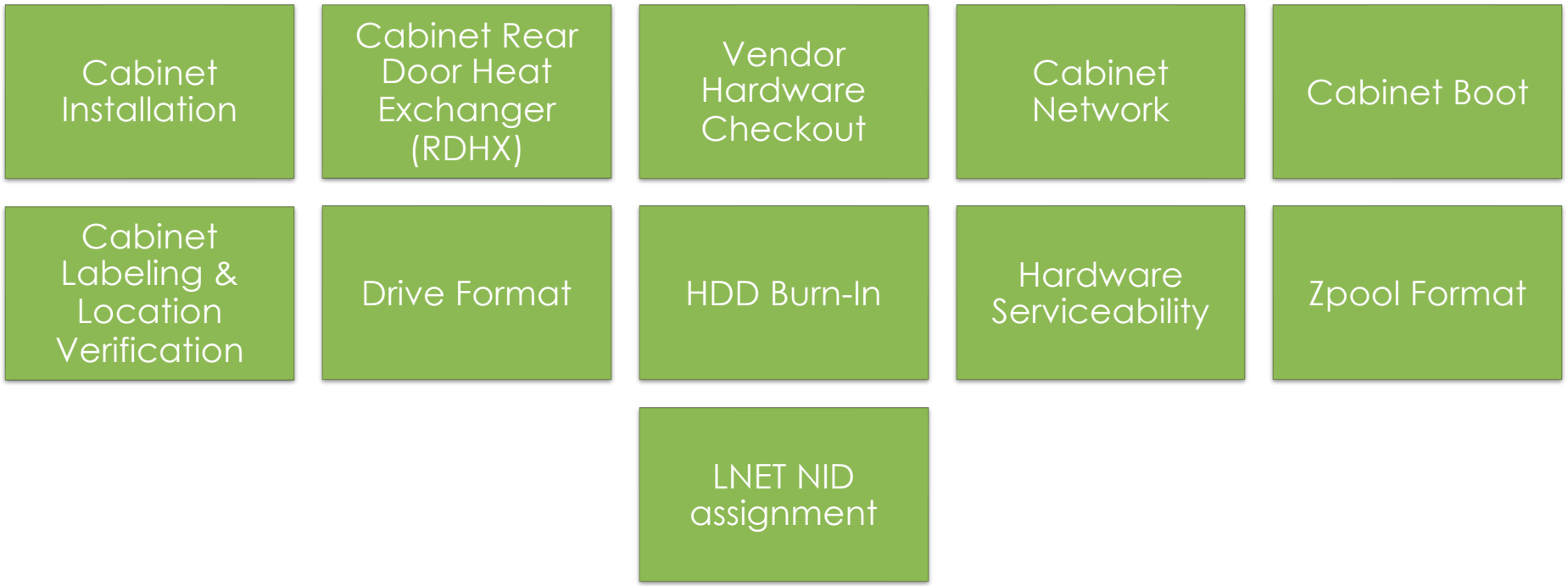
Hardware Test
Physical testing

Functionality Test
Demonstrate basic functionality meets resiliency, reliability and operational needs

Performance Test
Measure of hardware/software performance requirements

Stability Test
Verification that the storage cluster can withstand a workload similar to operational conditions

Acceptance – Phase 1 – Hardware
System install and checkout



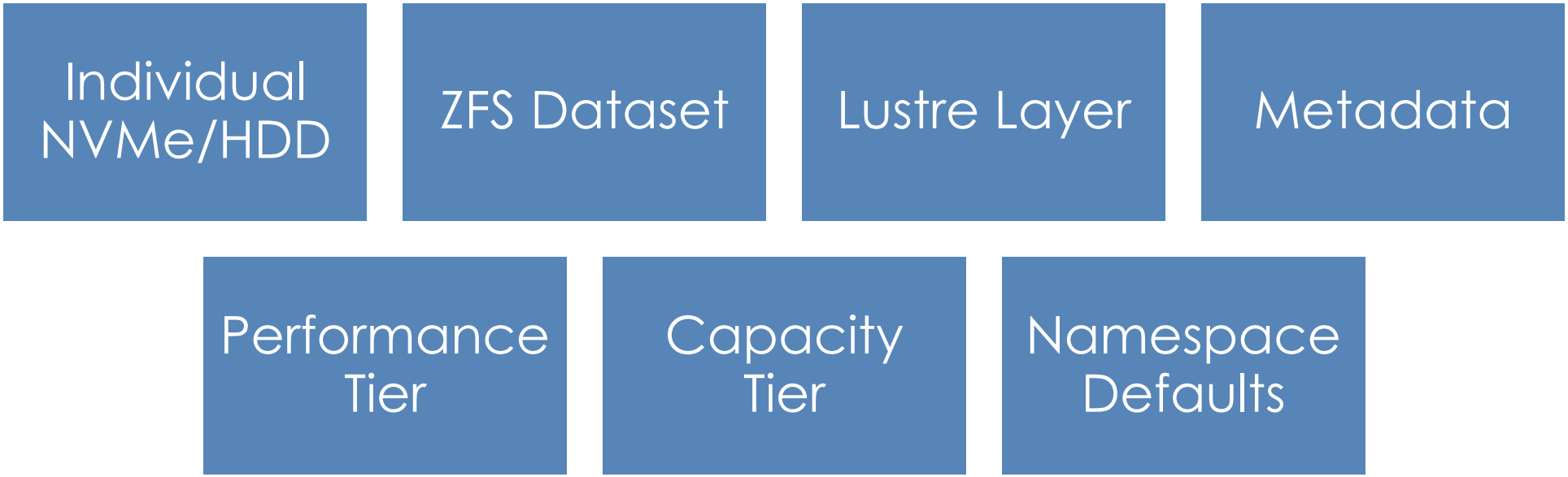
Acceptance – Phase 2 – Hardware
Single-unit testing

- Power Distribution Units (PDUs) monitoring
- Remote Power Control
- RDHX Health Monitoring
- System Monitoring
- Site Cybersecurity Integration
- Power Feed Resiliency
- SAS Network Resiliency
- High-Speed Network (HSN) Resiliency
- Serviceability and Cable Management

Acceptance – Phase 2 – Functionality
Single-unit testing

Power Cycle Resiliency	Location Beacon Test	Command-line Interface (CLI) Firmware Upgrades	Disk & NVMe Replacement/Fault Injection
Disk & NVMe Rebuild/Rebalance	Disk Variability	ZFS Parity Check on Read	High Availability Stack Testing

Acceptance – Phase 2 – Performance
Single-unit testing



Acceptance – Phase 3
Scale up

Scale-Up activities

Performance & system characteristics across various:

Access Patterns (File-per-process/Single-shared file)	Layout (unique/shared directory)	Tiers (DoM, Performance, Capacity, Namespace Defaults)	Block/transfer sizes	Read/write patterns	Degraded and pristine system health conditions	Artificially aged namespace configs

Acceptance – Phase 4
Full system testing

Functionality

- Namespace LFSCK
- System Health & Performance Monitoring
- Simulated hardware/software failures and maintenance activities
- Image management and deployment
- Tests from previous phases

Performance

- LNET Selftest
- "Hero" workloads (MDtest, IOR, …)
- Performance under simulated health issues
- Tests from previous phases

Stability

- Known I/O pattern
- Additional traffic from non-synthetic workloads
- Treated as-if the system is in full production with user workloads

Summary

Orion is in production and actively used

- Several users have reported significant I/O speed-up

Using PFL to provide a default layout that works well for many use cases

- No problems so far with DNE, PFL, DoM, etc.
- SEL provides protection against OSTs getting full

Acceptance process ensures the storage system is ready for end users

- Extensive process covers anticipated workloads
- Allows for a firm understanding of system behavior and limits

OAK RIDGE National Laboratory

Acknowledgements

This material is based upon work supported by the U.S. Department of Energy, Office of Science, Advanced Scientific Computing Research Program. 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.

OAK RIDGE National Laboratory

Questions?