OAK RIDGE National Laboratory

The Orion File System: Configuration and Performance

Rick Mohr (mohrrf@ornl.gov)

Dustin Leverman (leverman@ornl.gov)

Jesse Hanley (hanleyja@ornl.gov)

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

U.S. DEPARTMENT OF ENERGY

Introduction

- Orion is the latest center-wide Lustre file system deployed at OLCF and also the primary storage system for Frontier, the nation's first exascale supercomputer.
- Orion uses HPE Cray's ClusterStor E1000 storage platform

In this talk:

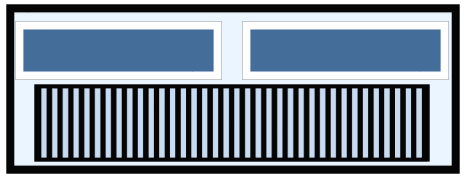
- Hardware overview
- Lustre file system
- File system performance benchmarks

Orion Hardware Overview

- Orion Lustre file system consists of:
  - 2 MGS nodes
  - 40 MDS nodes → 1x NVMe MDT
  - 450 OSS nodes → 1x NVMe OST, 2x HDD OSTs
  - 169 LNet router nodes
  - 12 utility nodes
  - 2 management nodes
- Capacity
  - 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)

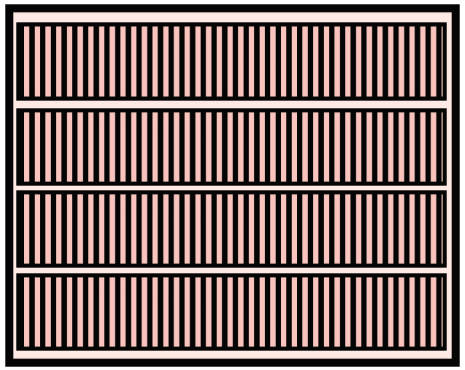
Orion Storage Enclosures

Gazelle



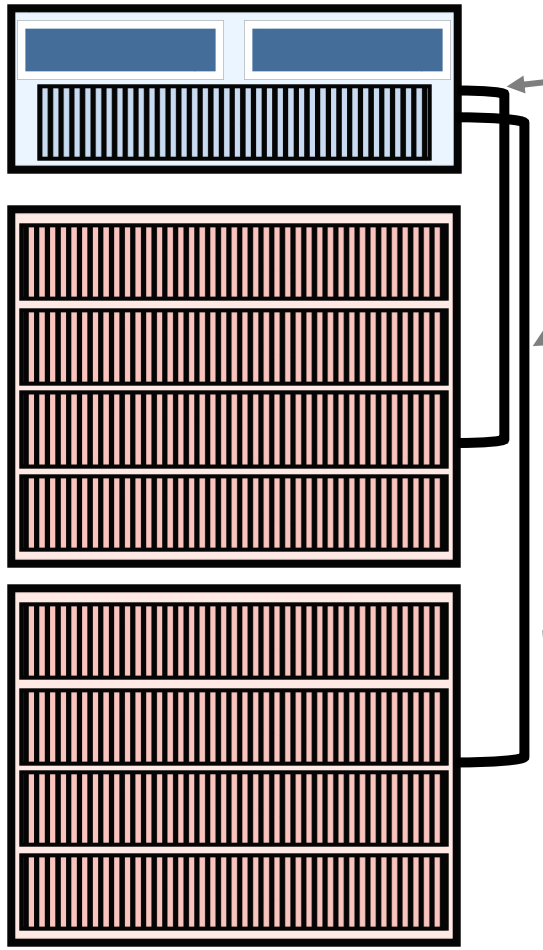
2 controllers, 24 NVMe drives

Moose



106 HDDs

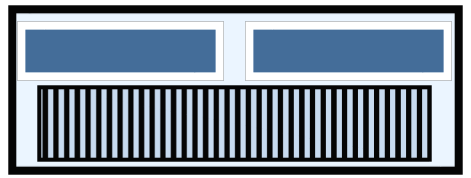
225x SSU (2x OSS)



24x 3.2 TB Samsung NVMe drives

106x 18 TB Seagate PMR drives

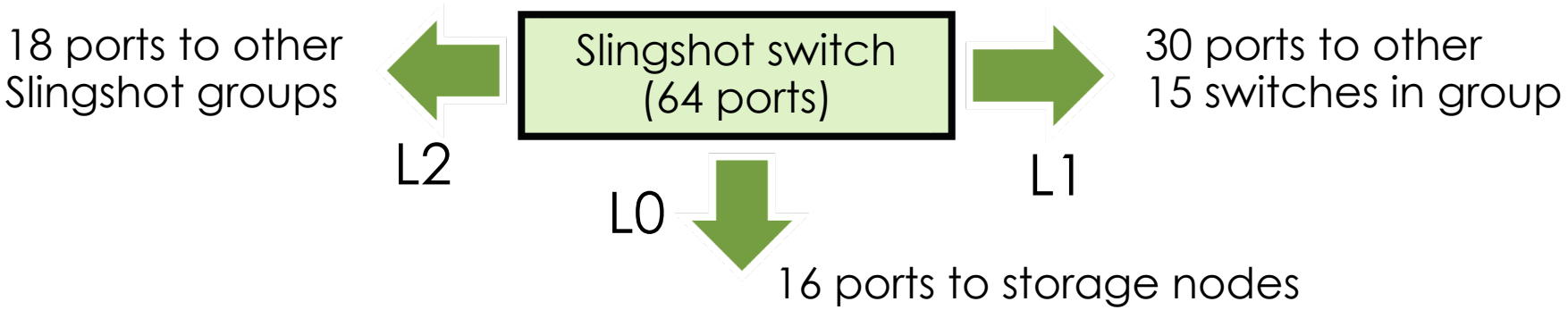
20x MDU (2x MDS)



24x 30 TB KIOXIA NVMe drives

High Performance Network

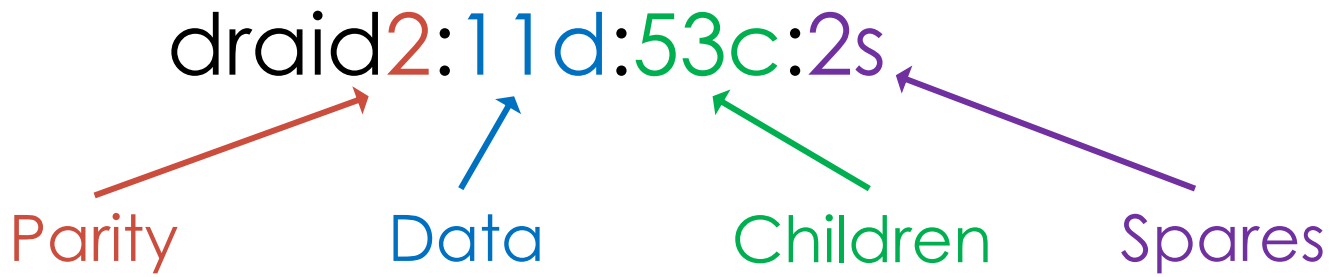
- Orion is connected to Frontier via Slingshot network
  - Dragonfly topology (switch groups connected by global links)
  - 16 host ports per Slingshot switch
  - Switches within group connected all-to-all
  - Remaining ports used to connect switches between groups
    - 50 GB/s between each storage and compute group



ZFS and dRAID

- Backend file system for MDTs/OSTs is ZFS v2.1.7
- Redundancy is handled using ZFS dRAID
  - MDT (NVMe) = draid2:9d:12c:1s
  - OST (NVMe) = draid2:9d:12c:1s
  - OST (HDD) = draid2:11d:53c:2s

All drives within same enclosure



Lustre Setup

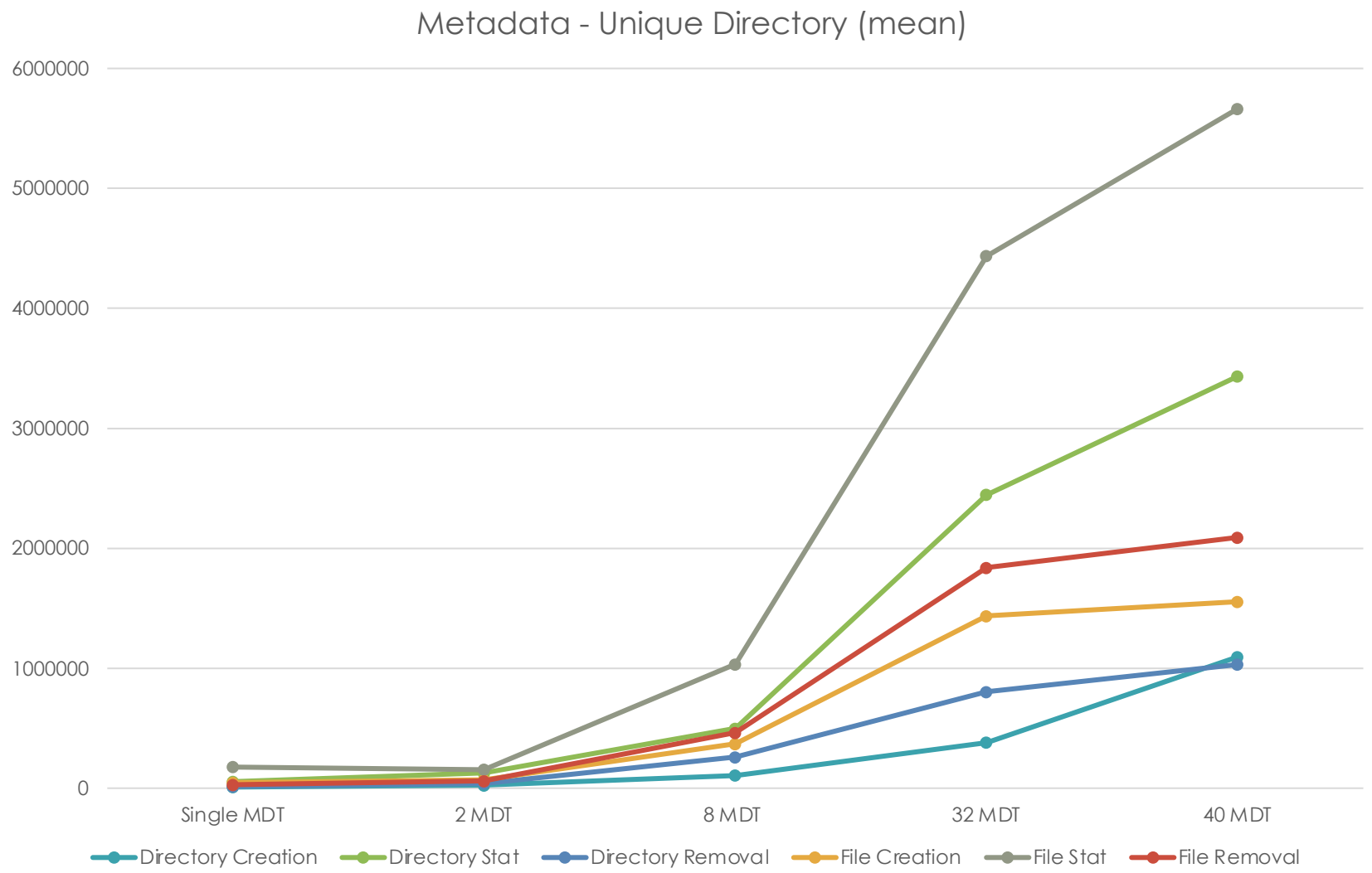
- Lustre version 2.15 w/ vendor patches
- Utilize two OST pools for file placement
  - ”performance” for all NVMe OSTs
  - “capacity” for all HDD OSTs
- Distributed Namespace (DNE) used to spread project directories across all MDTs
  - Only utilizing remote directories (DNE1) at this point
  - No striped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

Default File Layout

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

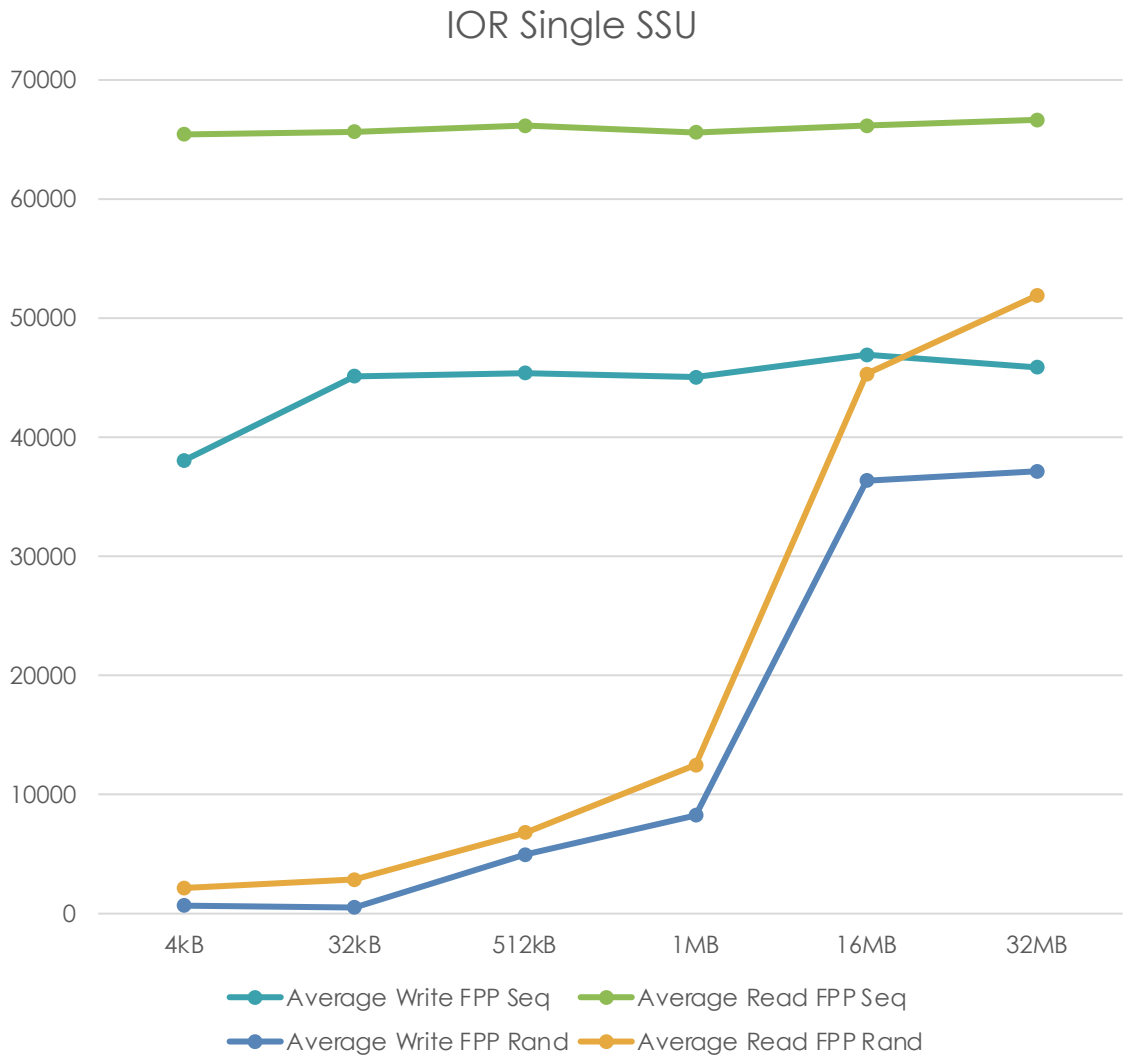
- Based on data from files on Summit, we expect:
  - 70% reside entirely on DoM
  - 18% reside on performance tier
  - 12% span to capacity tier (but will account for 99% of used space)

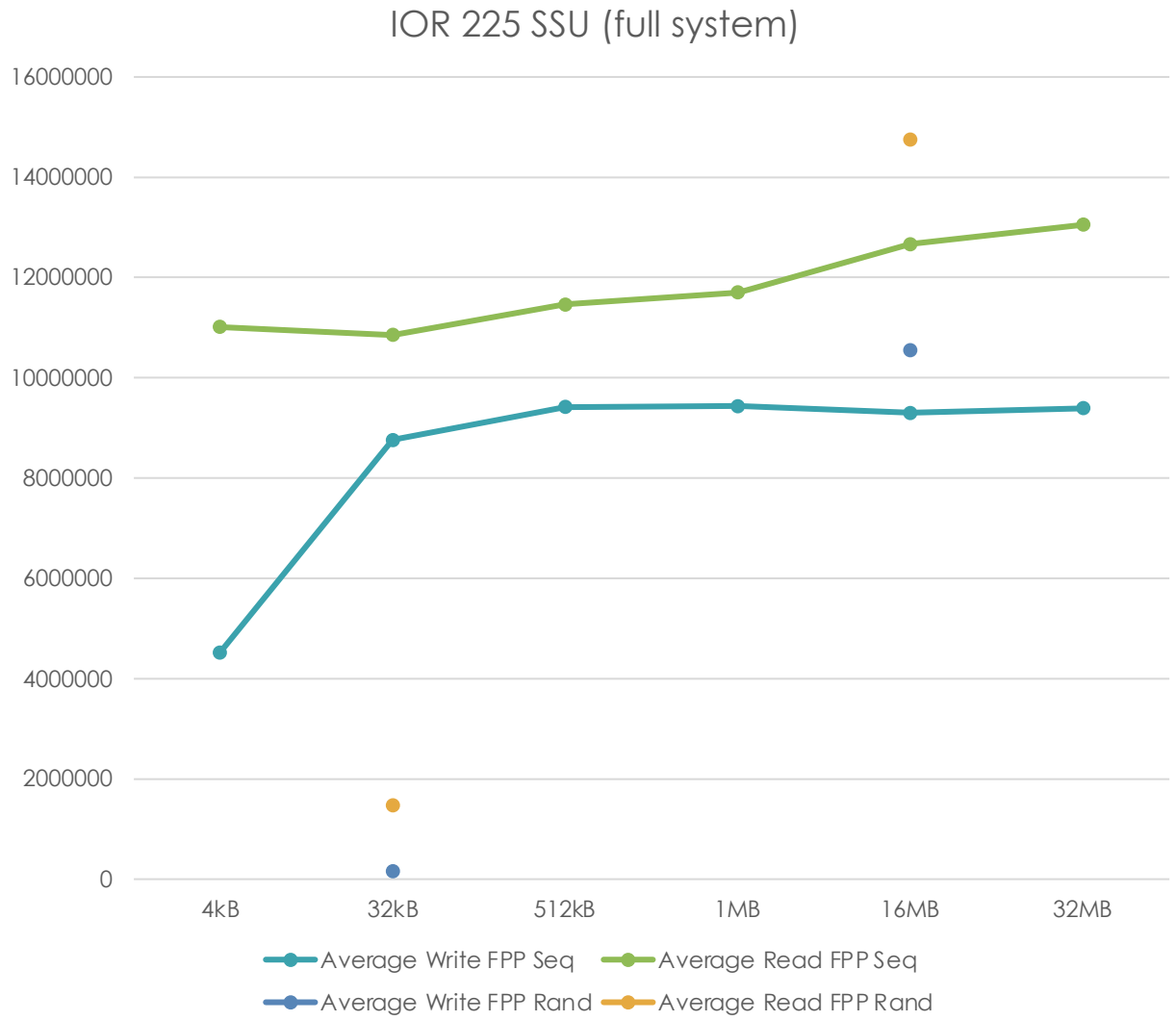
Orion MDT Benchmarks



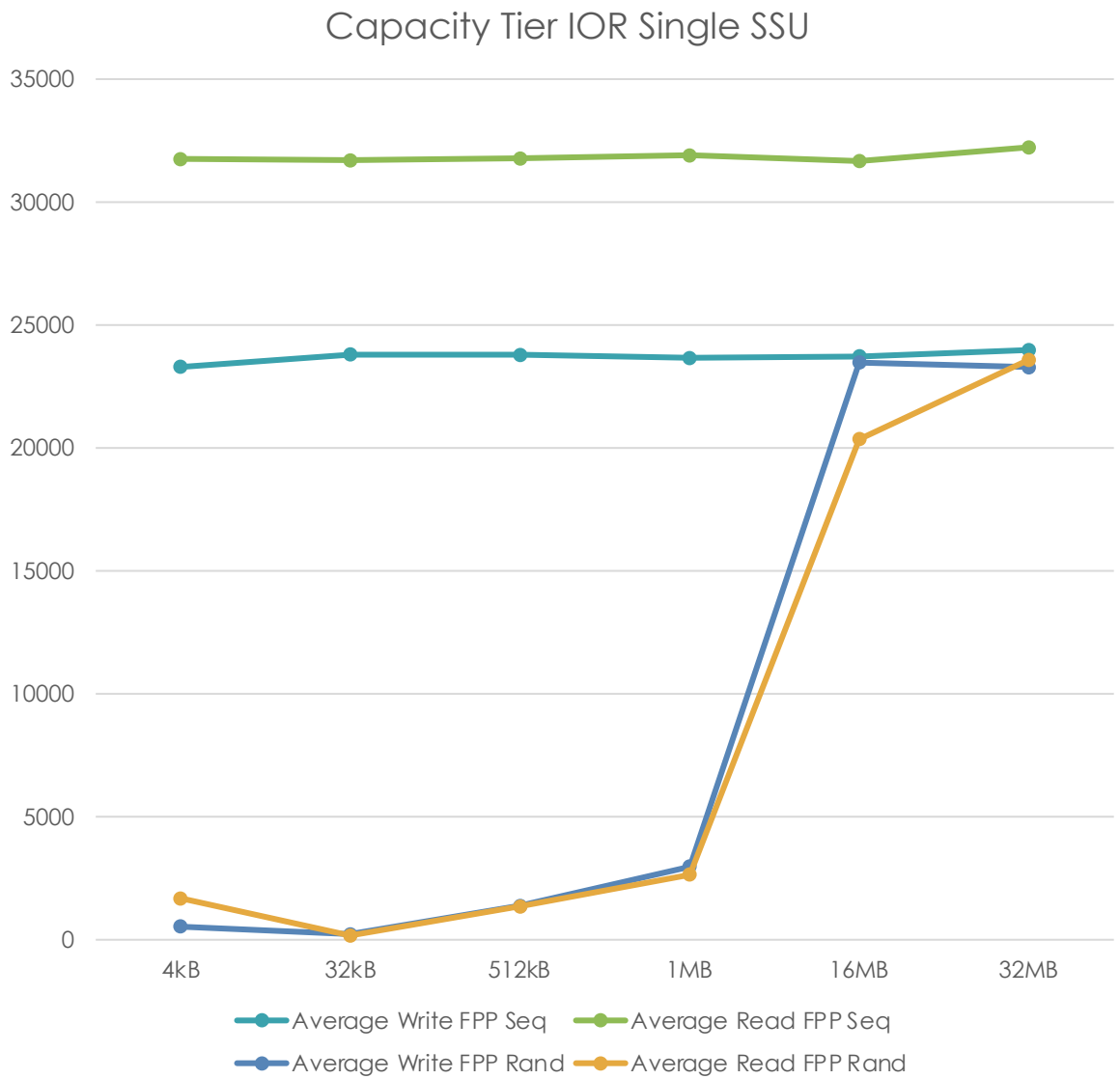
Metadata - Shared Directory (mean)	
	50% compute nodes
Directory Creation	17091
Directory Stat	70566
Directory Removal	20477
File Creation	25746
File Stat	98051
File Removal	23492

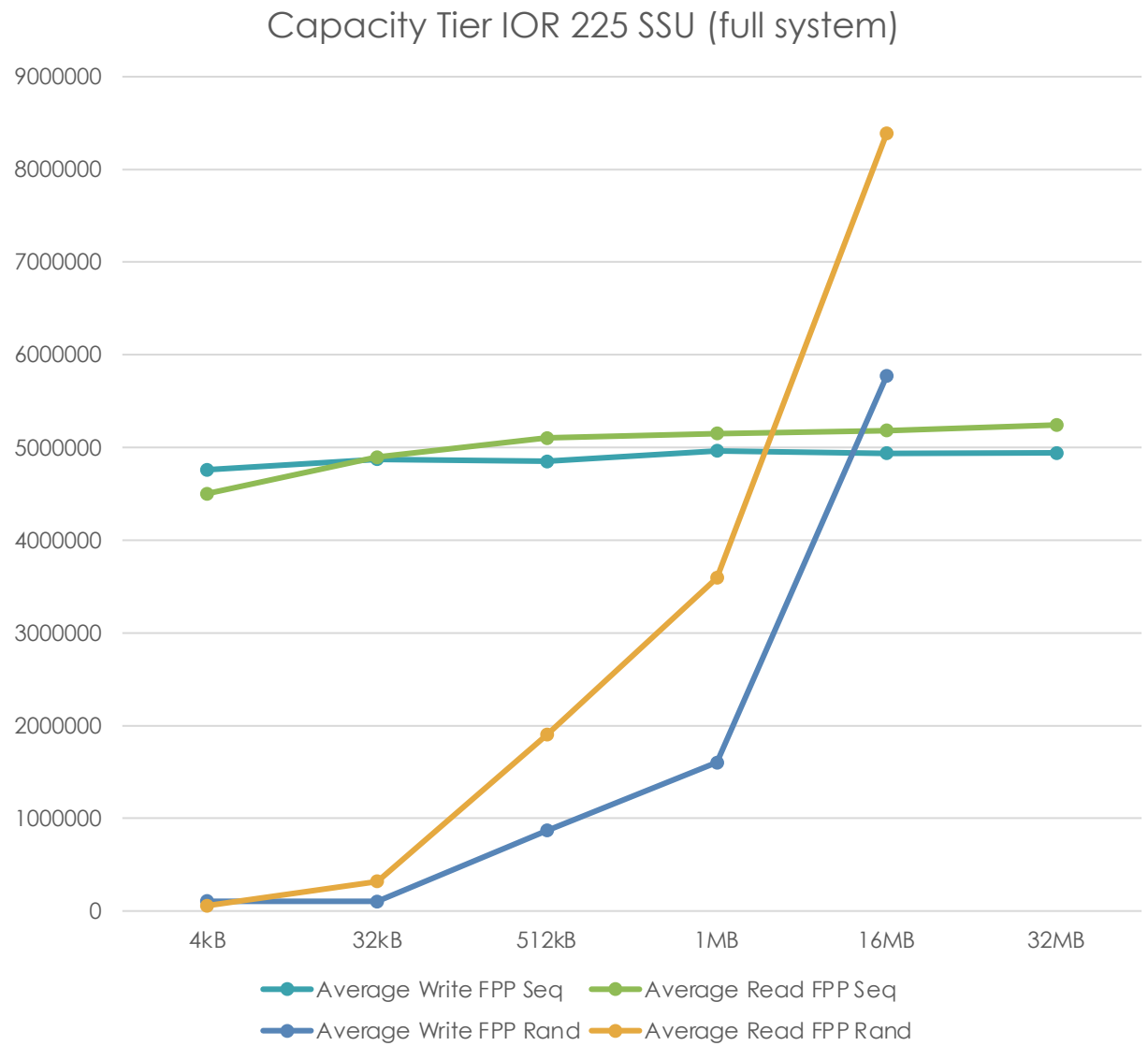
Orion Performance Tier Benchmarks





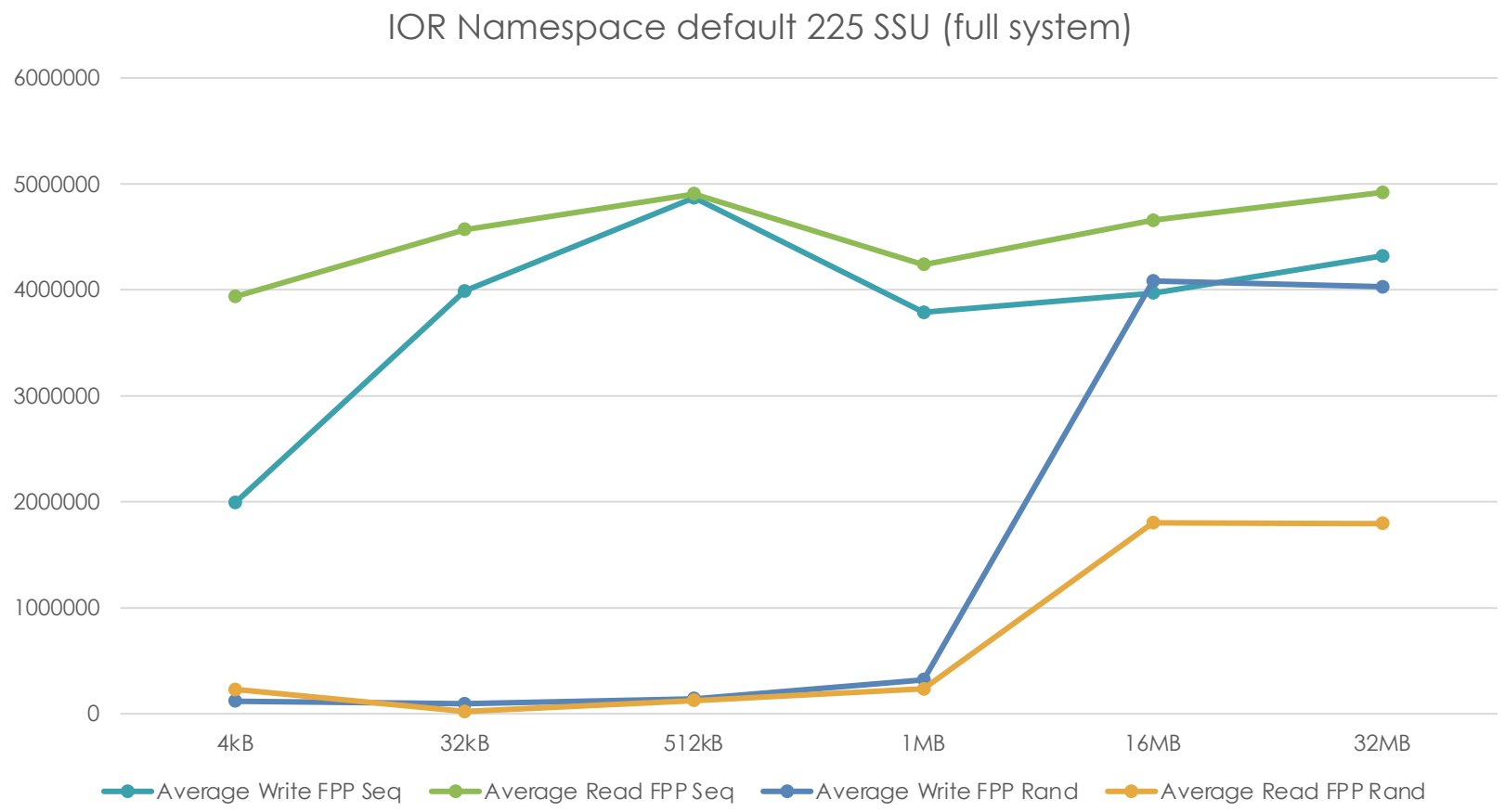
Orion Capacity Tier Benchmarks





Default File Layout Performance

Default layout: lfs setstripe -E 256K -L mdt -E 8M -c 1 -S 1M -p performance -E 128G -c 1 -S 1M -p capacity -E -1 -z 256G -c 8 -S 1M -p capacity "${TDIR}"



Summary

- Orion is in production and being actively used
  - Several users have reported significant IO 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
- Still some on-going testing and improvements
  - Network stability and fabric manager improvements
  - Working on policy engine testing for purging and file migration

Acknowledgments

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?