OAK RIDGE National Laboratory

Deep Dive into OLCF Storage Systems

ATPESC 2022 - Track 3 - I/O

August 5, 2022

Michael J. Brim, 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

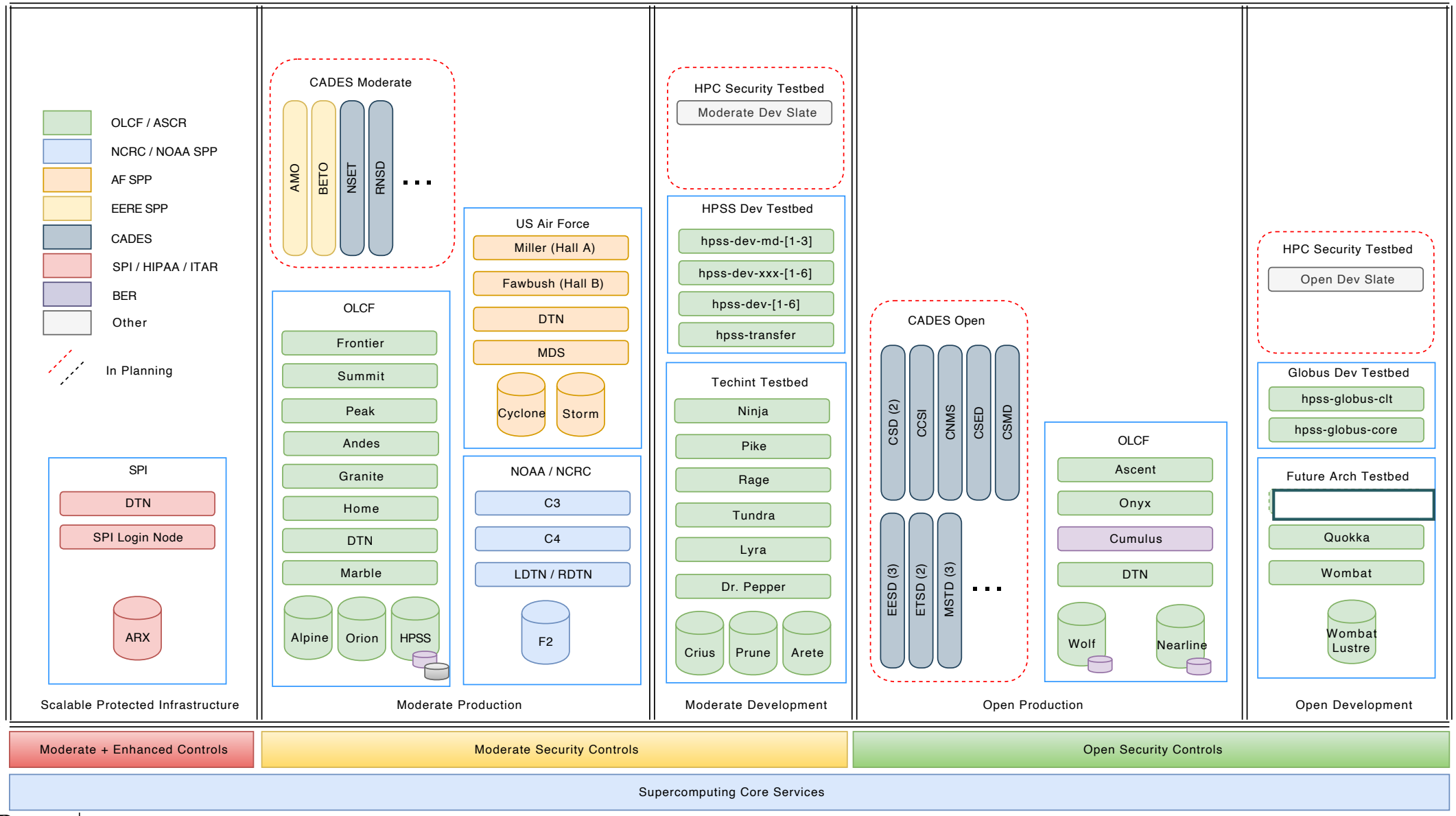
1. Quick background on NCCS and OLCF approach to HPC storage
2. High-level overview of Alpine - OLCF's current center-wide shared storage system
3. Deep dive on Orion - OLCF's next center-wide shared storage system

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	AFW	NCRC
Faster ->	**Job-term (24 hours or less)**	Summit Frontier			
	Short-term (less than 90 days)	Alpine 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/Nearline 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

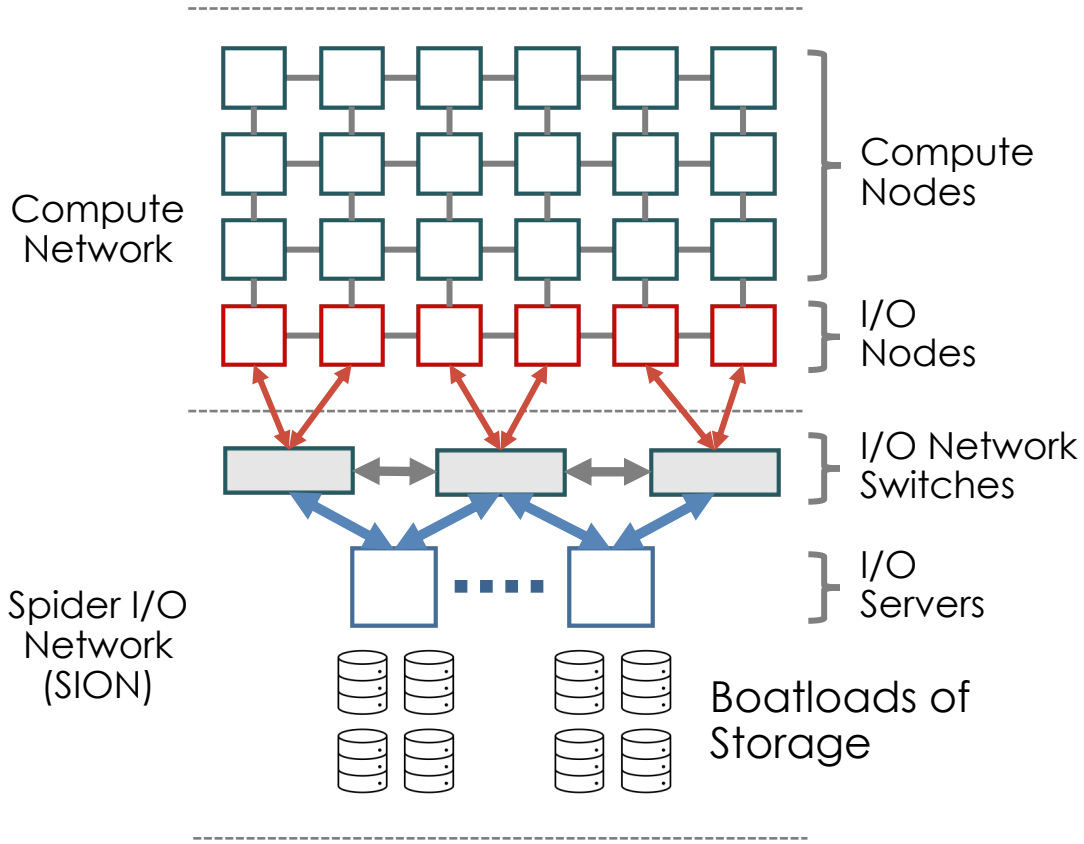
The OLCF Alpine Center-wide Shared File System

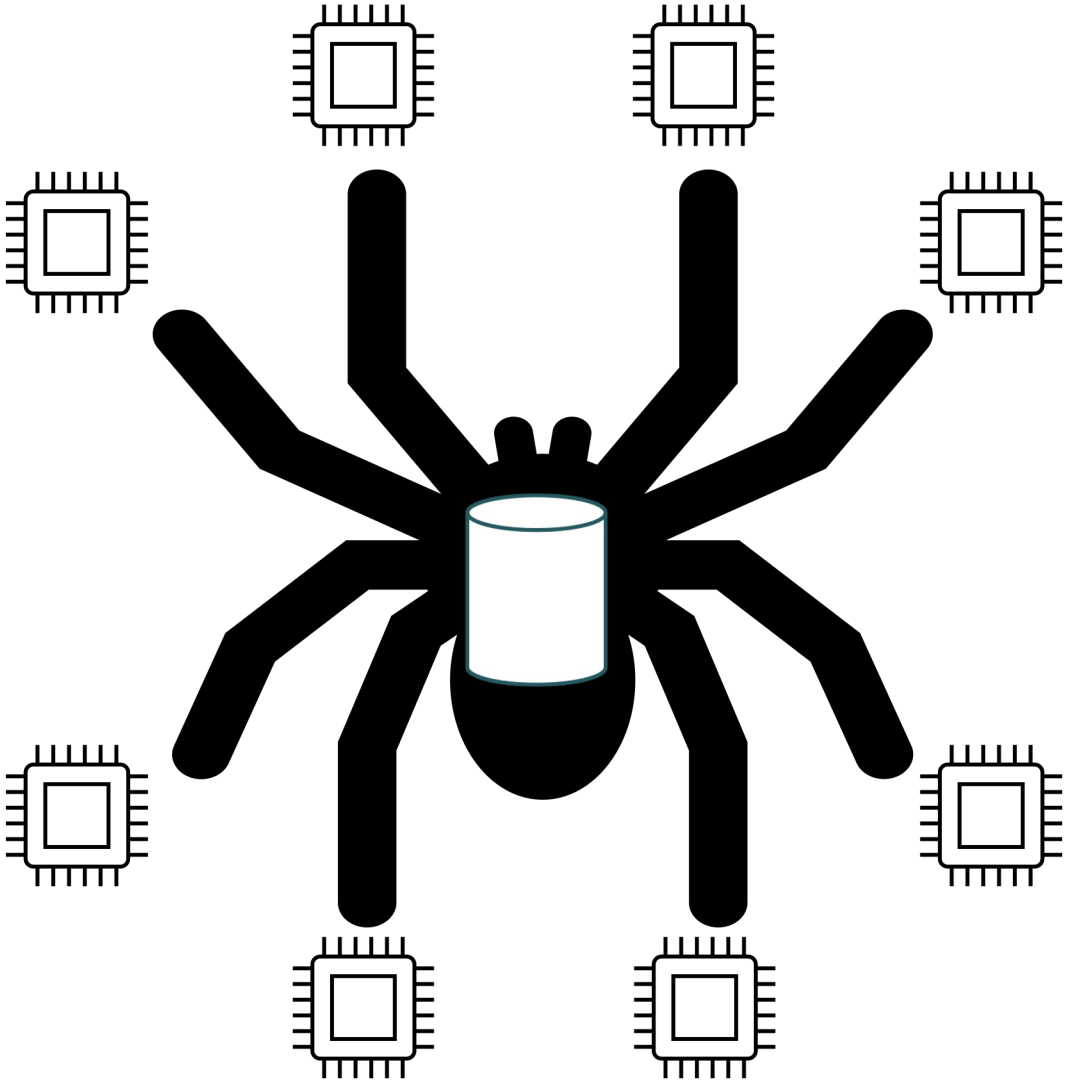
- Why center-wide?
- OLCF Alpine Details

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
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

Spider3 – Alpine

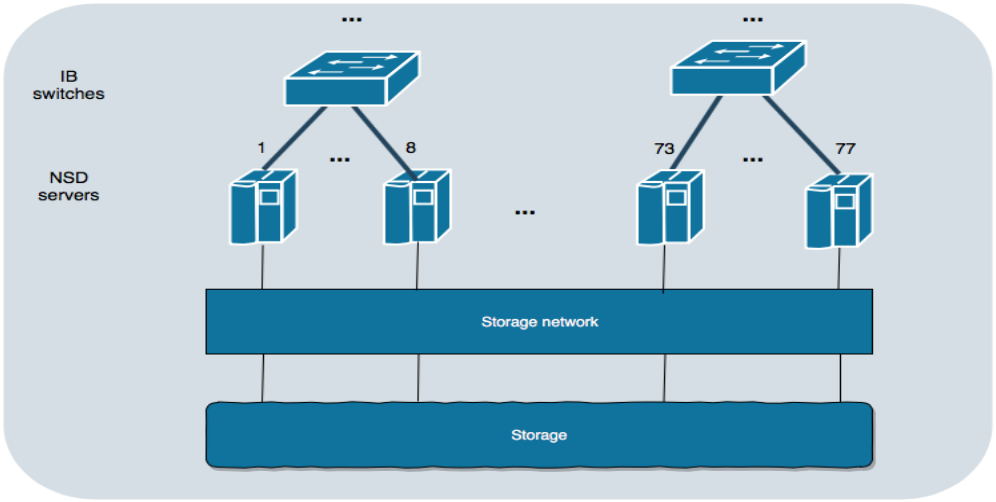
- 250 PB (¼ EB) usable capacity
  - 16 MiB block size
  - 16 KiB minimum fragment

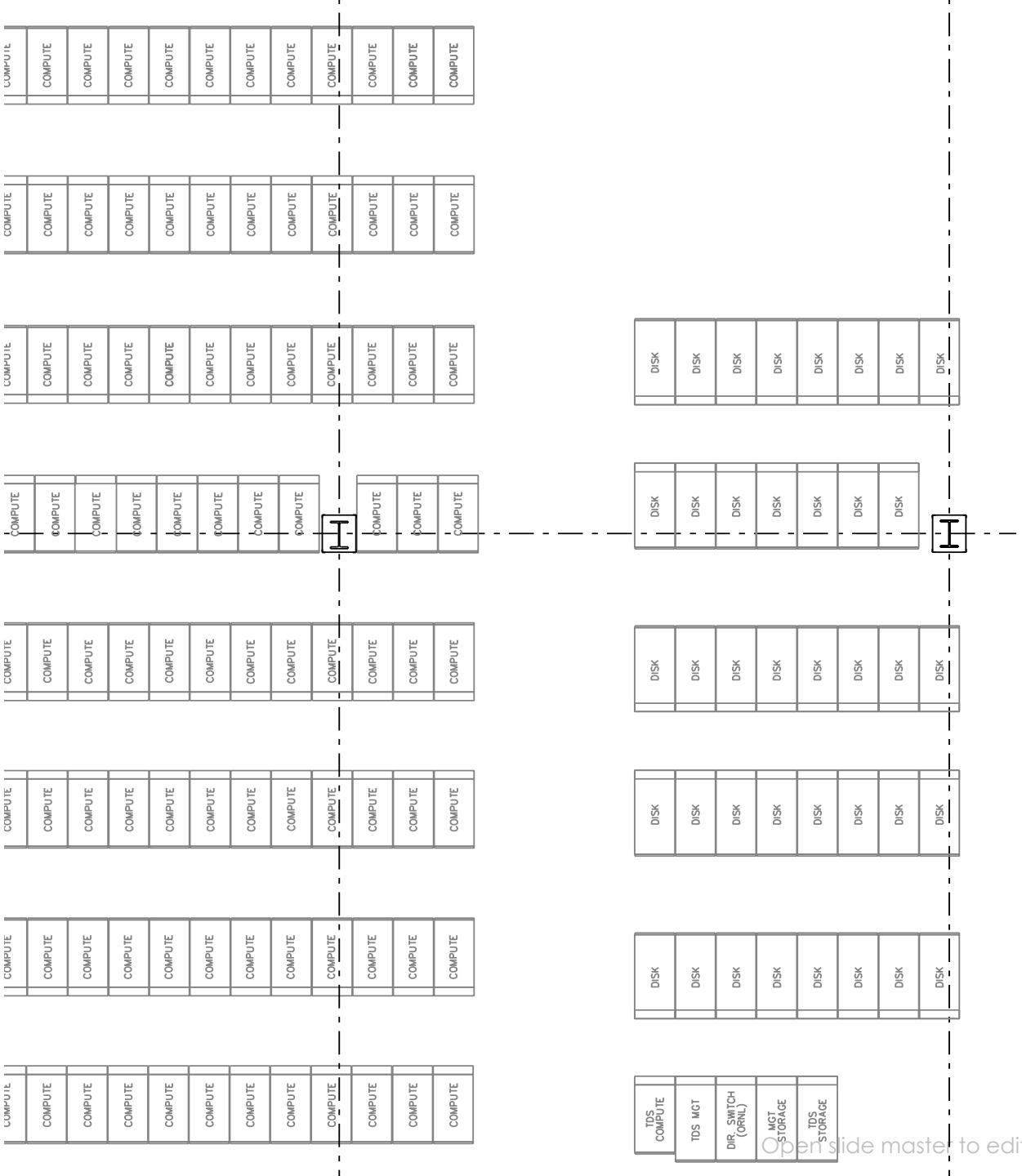
- 2.5 TB/s sequential read/write
- 2.2 TB/s random read/write
- 2.6M IOPs at 32KiB file size

- GPFS-5.x based
- GPFS Native Raid (GNR): de-clustered RAID with dual parity

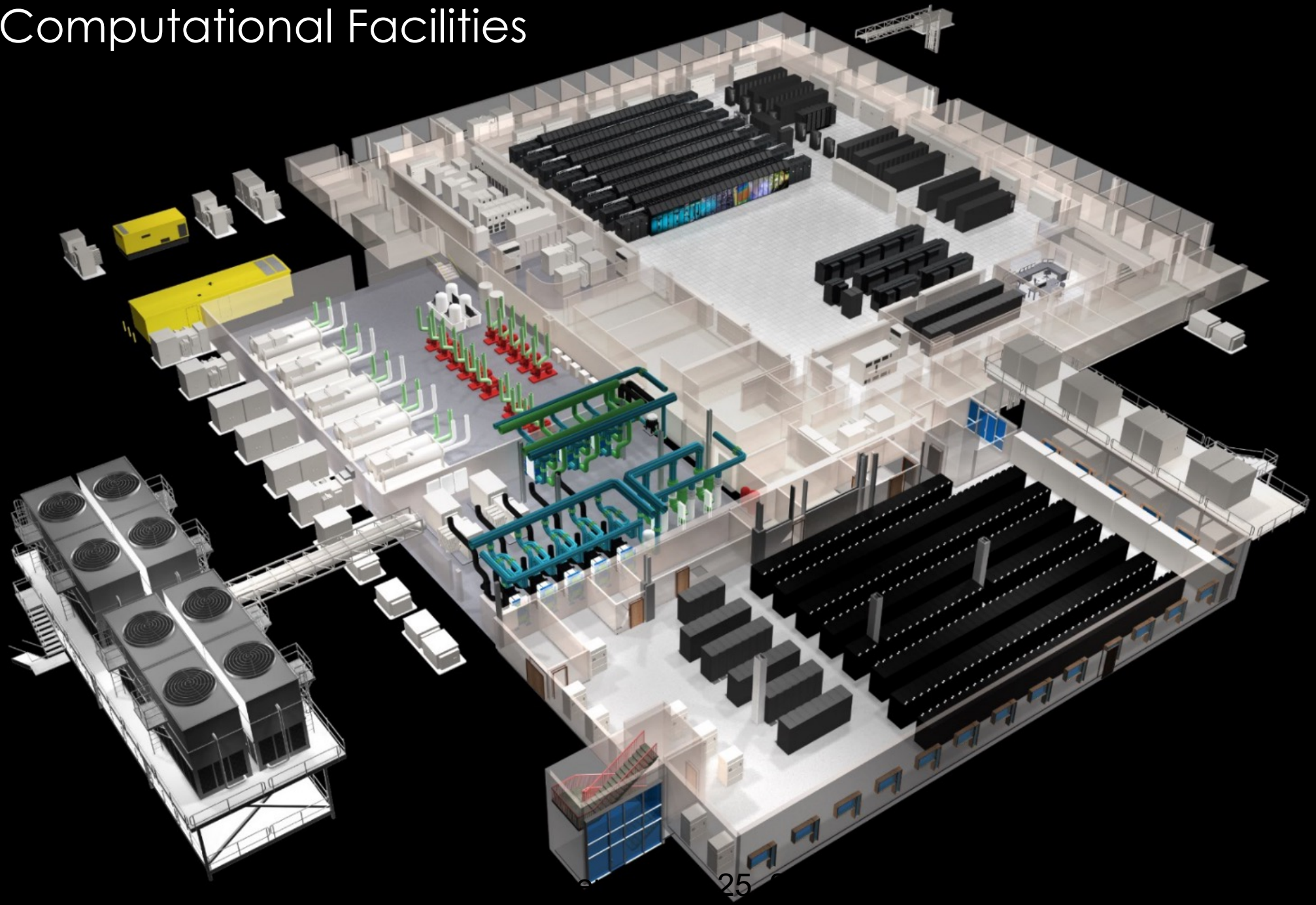
Alpine Layout

- 77+1x IBM Elastic Storage Server (ESS) GL4s
- Each ESS has:
  - 2x Power9 CPU
  - 4x EDR IB (total 100Gbps)
- 32,494 10TB NL-SAS drives





OLCF Computational Facilities



Spider3 Innovations/Improvements

- GPFS improvements due to OLCF contractual requirements:
  - Largest GPFS file system install (single namespace of 250PB)
    - GPFS tool improvements (fsck, etc…)
  - 10M files in a single directory
  - Max file size equivalent to Summit aggregate system memory
  - Metadata creates of 50k/sec in a single directory
  - RPC and GPFS daemon traffic performance improvements
  - Monitoring/metrics changes (big data store)

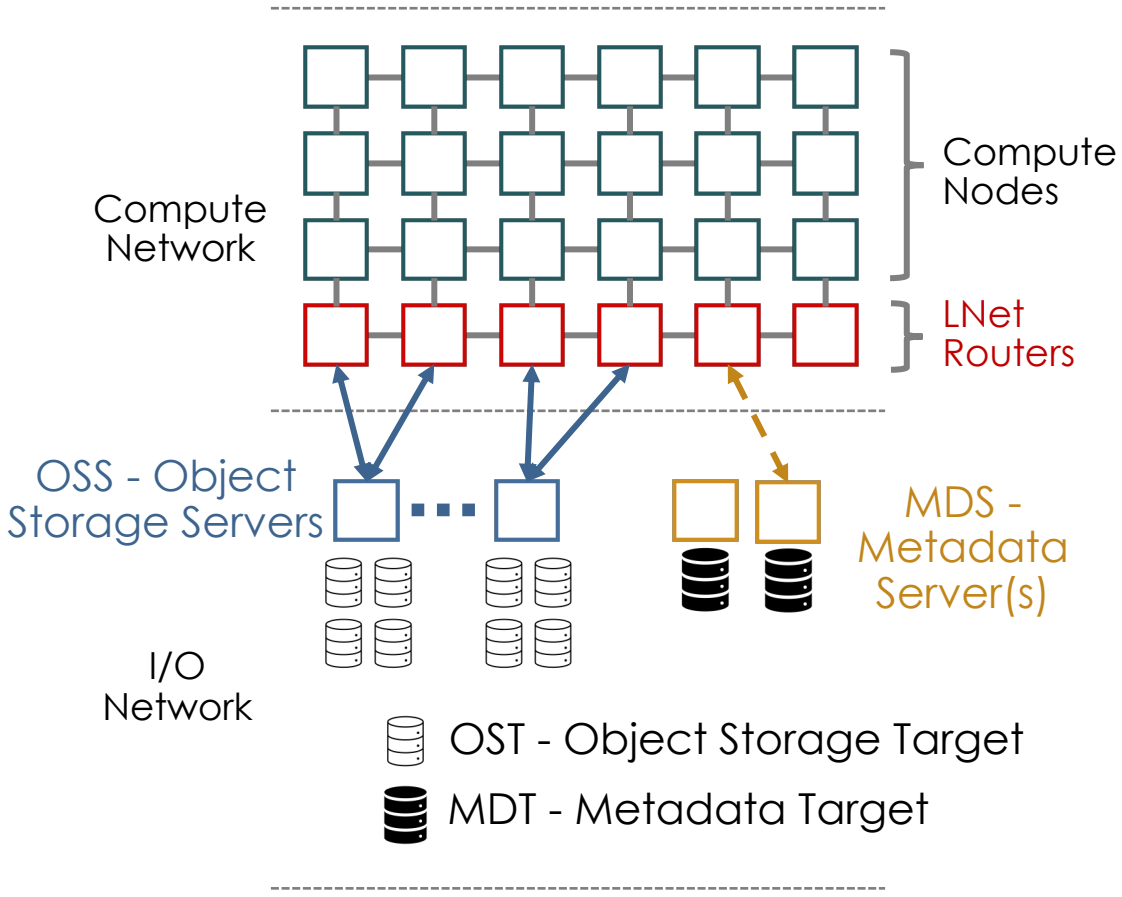
OAK RIDGE National Laboratory

OLCF Orion Center-wide Shared FS

- Architecture and Features of Lustre
- OLCF Orion Details

Lustre File System Architecture

- Two Types of Servers
  - MDS: maintain file system hierarchy, serve file metadata
  - OSS: serve file data
- Clients
  - talk to MDS to navigate FS, retrieve stats, and locate file extents
  - talk directly to OSS to read/write extents



Lustre File Data Management

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

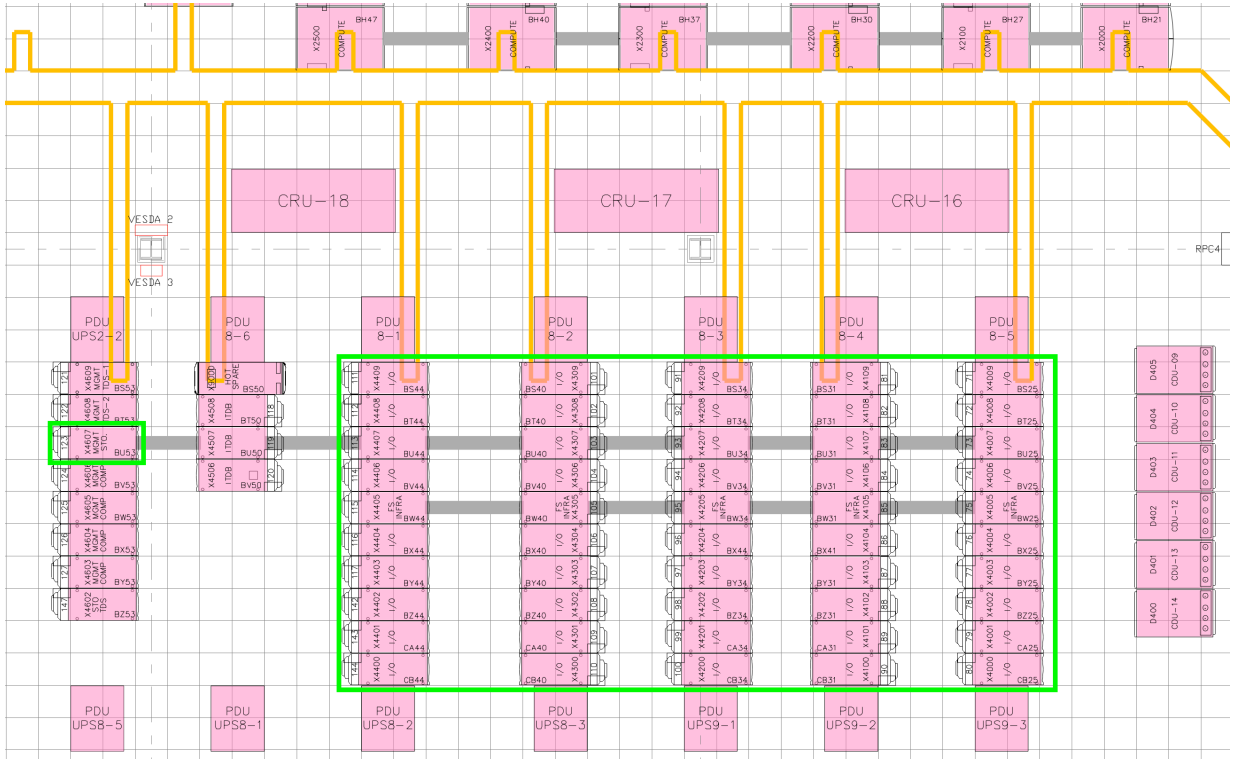
- Data on MDT (DoM)
  - for very small files, store the file data on MDT with its metadata
- Distributed Namespace Extension (DNE)
  - ability to employ more than one MDT to manage directories in a single file system (DNE1), or to stripe directory entries across MDTs (DNE2)
- Progressive File Layouts (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 Layouts
  - extension to PFL that avoids OST out-of-space conditions for small stripe sizes

Orion - Next-gen scratch filesystem

- Installed and acceptance ongoing
- 3 tiers:
  - Capacity:
    - 679 PB dRAID2:11d:2s
    - 5.5/4.6 TB/s R/W
    - 47,700 18 TB PMR HDD
  - Performance:
    - 11.5 PB dRAID2:9d:1s
    - 10 TB/s R/W
    - 5400 3.2 TB 3DWPD NVMe
  - Metadata:
    - 10 PB dRAID2:9d:1s
    - .8/.4 TB/s R/W
    - 480 30.7 TB 1DWPD NVMe

Orion Physical Attributes

- 40 MDS nodes (single MDT per node)
- 450 OSS nodes (2x HDD OSTs, and 1x NVMe OST per node)
- 160 Router nodes
  - route between Frontier Slingshot network and center-wide IB network
- 5x 16-switch Slingshot dragonfly groups



Orion Makes Extensive Use of New Lustre Features

- Performance tier to be treated as writeback cache implemented via policy engine
  - HPE-developed policy engine is plan of record
  - OLCF is implementing and testing another engine as a fall-back
- PFL will set striping policy to write to performance tier by default
  - SEL will protect us from ENOSPC in case ingest is too fast.

- Expected Striping Policy:
  - Files 512 KB-1MB will stay on DoM
  - Keep the next 1-2 MB on perf tier and never migrate to capacity tier
  - All file data over 2 MB will be eventually migrated to capacity tier

- Users should expect read performance from capacity tier

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