OAK RIDGE National Laboratory

April 22, 2026 | ALICE USA T2 Meeting

ALICE T2 at ORNL State and Directions

PRESENTED BY

Steve Moulton

Advanced Computer Science Research Section
Computer Science and Mathematics Division

U.S. DEPARTMENT OF ENERGY

ORNL IS MANAGED BY UT-BATTELLE LLC
FOR THE US DEPARTMENT OF ENERGY

ALICE ORNL T2 Instance Operations Team

- Managed at ORNL by
    - Steve Moulton
    - *Additional staff pending hiring decision*
- Networking support provided by ITSD (David Wantland)

ORNL Storage Element Capacity and Utilization

- Five FSTs in production, each having
  - 84 18 TB Drives each (ST18000NM004J or similar), raw capacity 7.5 PiB
  - RHEL 8.10 or 9.7

Instance	Size	Used	Files	Directories	PCR GB/TB*s	Use%	Vol-x	Path
aliceornleos	6.22 PiB	5.84 PiB	97.42 M	99.66 k	0.00	93%	1.00	/eos/aliceornleos

- Two legacy FSTs set aside for PRF
  - 60 8 TB Drives
- Two legacy FSTs idle for test environments
  1. 60 8 TB Drives
  2. 120 8 TB Drives (two jbods, one controller pair)

CCDB

```
EOS Console [root://localhost] |/eos/aliceornleos/> df /eos/aliceornleos/cond
```

Instance	Size	Used	Files	Directories	PCR GB/TB*s	Use%	Vol-x	Path
aliceornleos	6.22 PiB	9.08 TiB	97.44 M	99.66 k	0.00	0%	1.00	/eos/aliceornleos/cond

QuarkDB

Three instances in quorum

- ornl-eos-01 (mgm)
- ornl-eos-04 (to be backup mgm)
- eos-fst-12

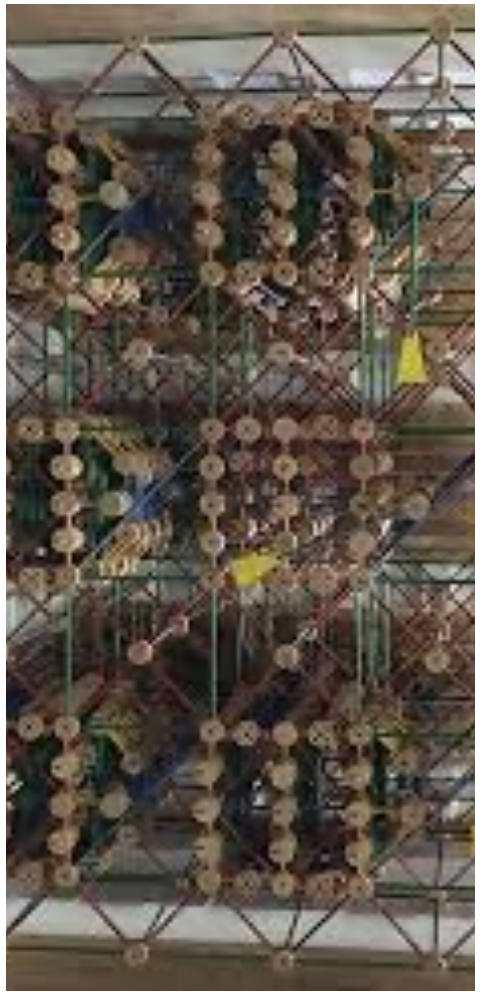
Data backed up via

- snapshots to zfs (itself snapshotted)
- S3

Raft check automated, and complains loudly if there is an issue (yellow or red)

CE Status

- 105 hosts in CE. All in quad chassis
  - 11 Phase 1 Dell 6220 still functioning. Will not boot RH10 kernel
    - Most power supplies scavenged for Phase 2 chassis
    - Superannuated (original CADES nodes ca 2014) – no avx2
  - 54 Phase 2 Dell 6220-II 2*8*2 Cores 64 GB – no avx2
  - 8 Phase 3 Cray Intel S2600KPR 2*18*2 Cores 256 GB
  - 4 Phase 4 Supermicro SYS-2029BT-HTR 2*24*2 384 GB
  - 28 Phase 5 Supermicro SYS-620TP-HTTR 2*24*2 256 GB
- Consistently running 105 jobs using 5376 HT cores (or 672 8-core job slots)
- Worker nodes all running Rocky 9.7. Automated/identical configuration.
- Phase 1 & 2 hosts running on borrowed time.



Support Services on alice-hyper03

Small ZFS instance (35TB) for cluster-wide shared file system needs

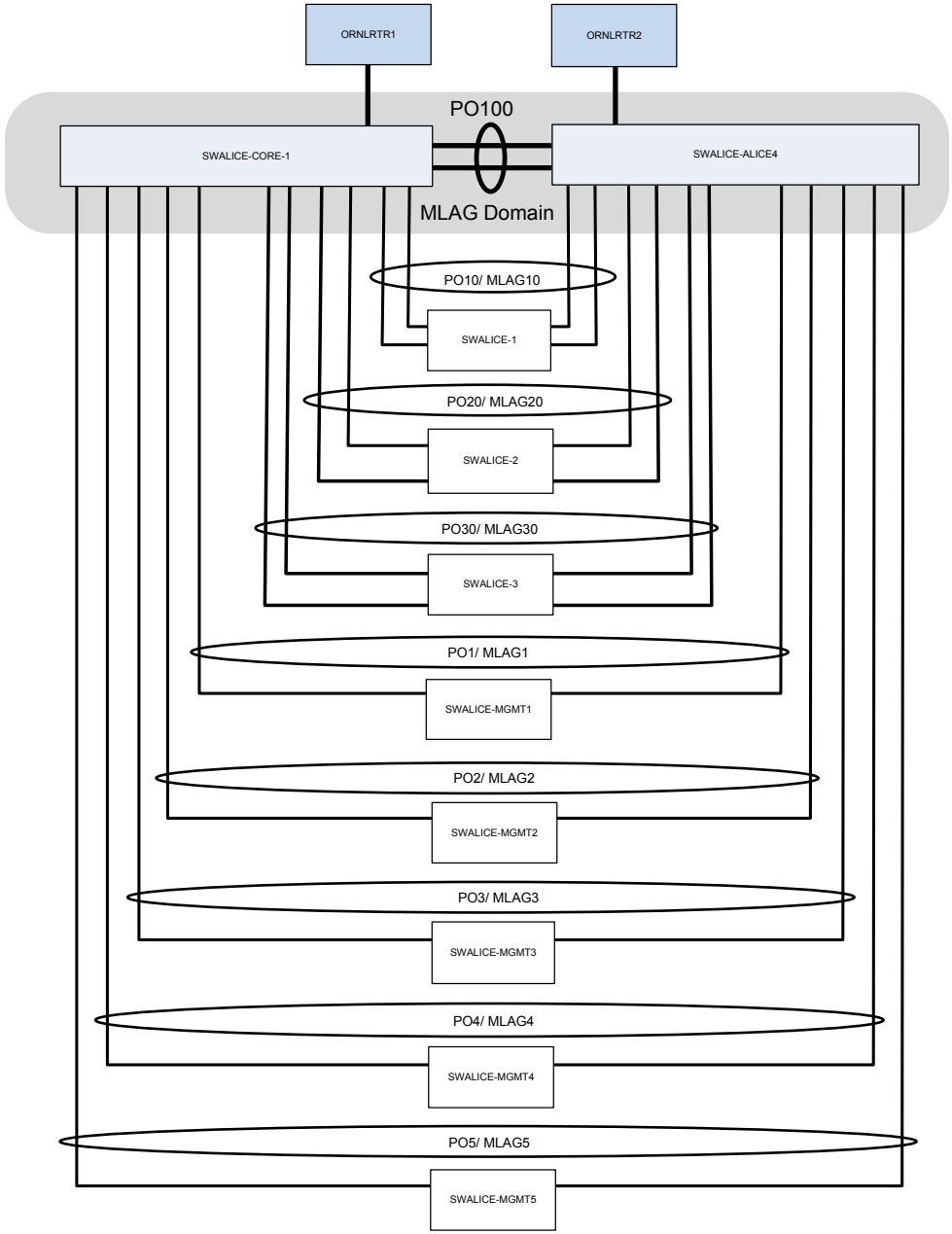
- /var/lib/condor_ce mounts here
- Operational and utility storage as needed.
- VM snapshots go here (virtnbdbackup). Previous two months daily/monthly snaps always available.

Virtual Machines

- Login
- Checkmk for operational monitoring and metric gathering
- Slurm instance (gets submissions from vobox)
- Vobox
- Squid (frontier-squid) for CVMFS
- Service (dnsmasq, email gateway, cluster trusted host, ansible, etc).

ORNL ALICE Network Topology

- Core switches (at top) have direct access to ORNL border routers.
- Two redundant paths to ESnet
  - 40Gb upgradable to 100Gb
- Older (phase 1, 2) nodes have 1Gb connectivity to switches.
- Newer nodes all have either 10Gb or 25Gb connections
- All production FSTs have a 25Gb connection
- Based on monitoring, none of the network connections are saturated



SE hardware reliability issue - resolved

FSTs 13, 14, 15 (Gigabyte S452-Z30-00 systems) all failed after firmware and bios upgrades. Systems bricked. Not even console presented.

- Short term solution: drain the BIOS
  - Short the BIOS reset pins for at least 15 minutes -- or --
  - Pull the CMOS battery for at least 15 minutes
  - Reboot then completes, but takes a LONG time (over an hour)
- VGA is never found, either at console port or in bmc html5 interface
- Vendor has replaced all backplanes, in one case twice . FSTs are now largely stable.
- Older backplanes on FST11 and FST12 were not affected

EOS Issue at ORNL – Resolved?

The MGM occasionally goes unresponsive. **`eos fs ls`** times out. This now happens two or three times a month. Before last EOS update this would occur several times a week.

- **`systemctl stop eos@mgm.service`** typically takes three to four minutes to complete
- **`systemctl start eos@mgm.service`** completes in normal time, and restores all functionality
- MGM deployed on RHEL 8.10. Will redeploy to RHEL 10 as time/staffing allows

EOS Issue at ORNL – Resolved? (2)

- Status monitoring now automated via script written at LBNL. ORNL instance checks MGM status every ten minutes via **`eos fs ls -d`**
- No occurrences between March 24 and (sigh) early this morning.
- This morning: multiple occurrences of a memory leak in libasan (system shared library) in /usr/bin/hostname at service invocation
    - Due to older OS release?

I do not intend to worry about it until the MGM is redeployed on Rocky/RH 10.

Security Posture Change

Scientific DMZ no longer a thing. ALICE instance moved to Open Research PZ.

ORNL configuration management and security software now mandated. Mild performance hit (nessus & elastic-agent)

Firewall posture has not changed

ORNL T2 Roadmap

- Complete IPv6 deployment
- Enable HTTPS for Xrootd
- Upgrade all systems to RHEL/Rocky 10.
    - Time frame dependent on ALICE & EOS
- Deploy legacy hardware as PRF EOS instance.
- Deploy second MGM as hot spare for ALICE instance
- Redeploy networking so all BMCs are on distinct VLAN
    - Ensure all systems are reachable via BMC (HTML5)
        - This does not include Cray nodes and older. Management via 

ORNL T2 Roadmap

- Monitoring redeployment and accessibility by non-ORNL staff
- Redeploy log file aggregator (Graylog)
    - Could do this in hypervisor, but this really keeps disks busy

In general, factor out ad-hoc changes and drift

Potential SE upgrade (new ornl-fst-16)

Supermicro Hyper Server 2115HS-TNR

- 32 core AMD EPYC 9355P
- 384 GB
- 2 * 240GB SATA SSDs
- 2 * 25G SFP28 (Mellanox CX-6 LX)

Seagate JBOD with 84 x 24tb drives

Description in quote incorrect. Quote is confirmed for 84x24 (2 PB).

Essentially the same as eos-fst-[11-15], except 24 PB disks and Supermicro-based

Budgetary quote is $107,519.59

Quote pending for 18TB disks.

Potential CE upgrade: new quad

Supermicro SuperServer 221BT-HNTR BigTwin - 4 DP Nodes

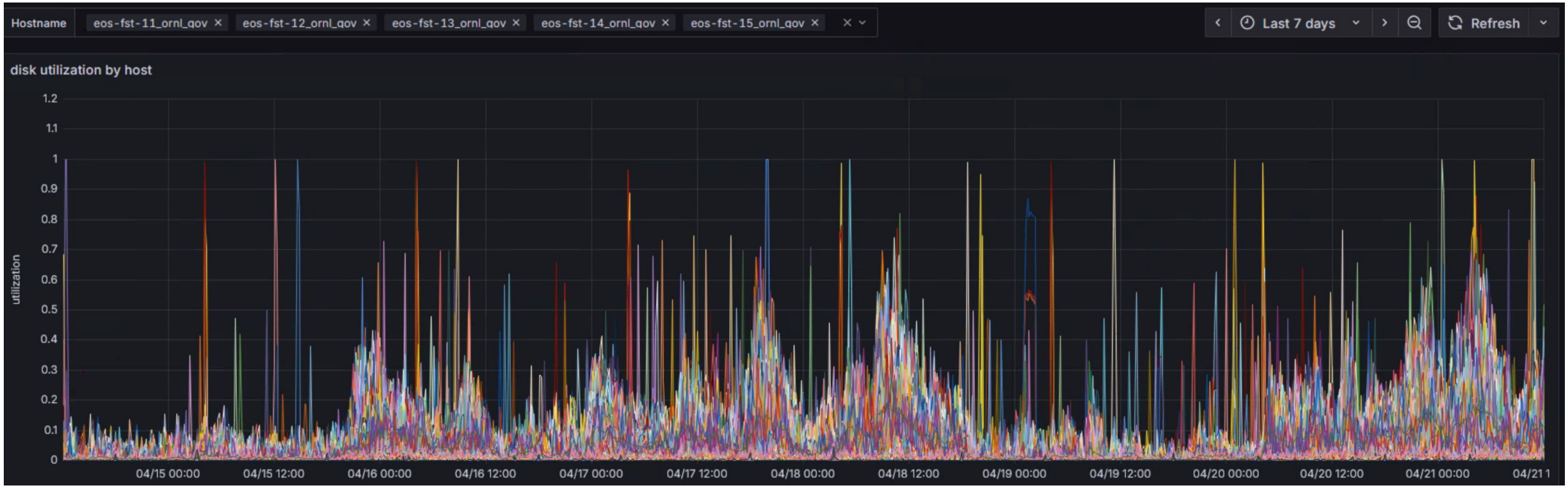
- 8 x Intel Xeon Gold 5520+ Processor 28-Core 2.2GHz 52.5MB Cache (205W) (2 per node)
- 32 x 32GB PC5-41600 5600MHz DDR5 ECC RDIMM (256 GB per node)
- 8 x 960GB Samsung PM9A3 Series U.2 PCIe 4.0 x4 NVMe Solid State Drive (7mm)
    - Could reduce size on system drives
- 4x 25GbE SFP28 (dual port)

$90,986.62

Maybe we get storage this year

Disk Saturation (from yesterday's discussion)

Shows all 420 FST disks



OAK RIDGE National Laboratory

U.S. DEPARTMENT OF ENERGY

ORNL IS MANAGED BY UT-BATTELLE LLC
FOR THE US DEPARTMENT OF ENERGY