OU XRootD Site Report

Horst Severini

XRootD and FTS Workshop 2024

September 2024
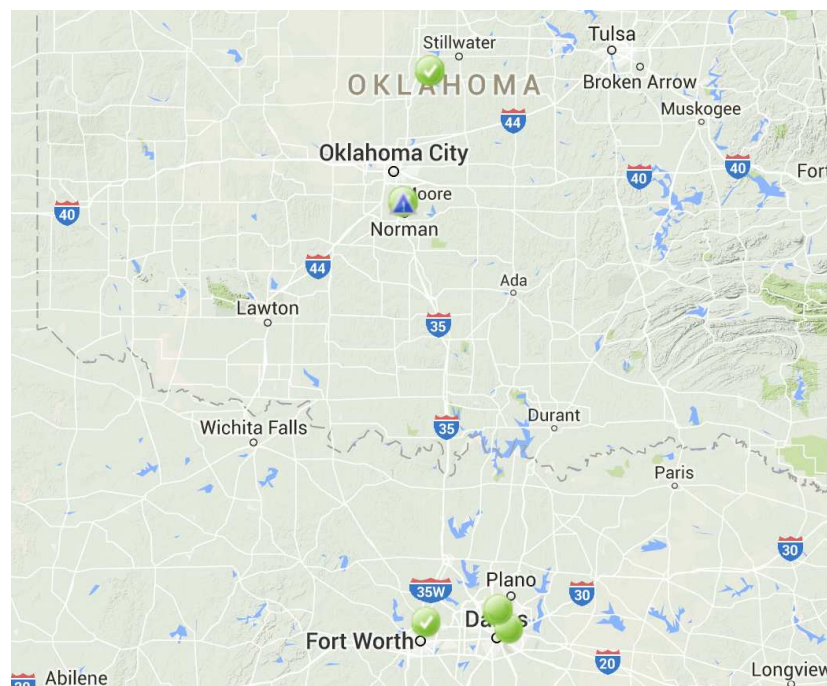
Outline

- Computing and Storage Hardware

- Network

- XRootD Configuration

- CephFS Testing and Migration Plans

US ATLAS SWT2 Center

- **University of Oklahoma**

    - OU Supercomputing Center for Education and Research (OSCER)

    - OU High Energy Physics Tier-3 cluster (OUHEP)

- **University of Texas Arlington**

    - Chemistry and Physics Building (CPB)

OU_OSCER_ATLAS Tier-2 Hardware

- 85 Nodes (5600 Slots) – 2 GB RAM per Slot

- 10 Support Nodes (1 GK, 1 DTN/SE, 1 XRootD redirector, 7 XRootD storage nodes)

- 700 TB of usable XRootD storage (7 Dell T630s with 16 8 TB drives, RAID6, xfs)

- SALT, SLURM 23.02.5, XRootD 5.4.2

  ○ GK running AlmaLinux 9

  ○ New SE also running AlmaLinux 9, XRootD 5.7.0

OU_OSCER_ATLAS Tier-2 Hardware

- Tier-2 hardware Part of generic OSCER HPC cluster

- Rest of OSCER Schooner Hardware

  ○ 850 Nodes (about 25k Cores) – 2-4 GB RAM per Core

  ○ Opportunistically available for ATLAS production

  ○ Have gotten up to an additional 5k cores when local demand was low

- SWT2_CPB

  ○ 548 Nodes (21296 Slots) – 2-3 GB RAM per Slot

  ○ 50 Support Nodes (12 head, 38 storage)

  ○ 13.1 PB of usable xrootd storage (MD3X60/R740XD2)

Network

- OU connected at 100 Gbps to OneNet, which is now 2x400 Gbps

- OSCER connected at 100 Gbps to Internet2 and ESnet via OneNet

- Working on CC* Proposal to upgrade OSCER core to 2x400 Gbps

- SE on 100 Gbps DMZ – OFFN (Oklahoma Friction Free Network)

- Current (old) SE:

  - Dual 25-gig Bonded

  - WAN XRootD transfers to storage nodes up to 25 Gbps

- New SE Dual Bonded 100 Gbps NIC

XRootD Configuration

- Currently, se1.oscer.ou.edu acts as XRootD server for 700 TB XRootD cluster

  - Pretty stable and performant

  - Occasionally, one or two of the 7 servers gets overloaded with open connections, causing transfer timeouts

  - Restart of XRootD service on these nodes eventually fixes this

Ceph Migration Plans

- Plan to migrate to CephFS file system after storage server warranty expires in November 2024

  - OU Research Disk – OURdisk

  - 9.5 PB and growing

  - 14 GB/s total througput

  - Very performant and secure, reasonably priced

  - $93 per usable TB, good for 7 years

OU Ceph Setup

- About 35 Dell R740xd2 storage nodes

- 24 18 TB HDDs per node

- 8+3 erasure coding at the server level:

  - 8 data chunks + 3 redundancy chunks

  - Monte Carlo simulation of Ceph was unable to run enough randomly generated realizations to induce a single instance of 4 simultaneous HDD failures, on 1000+ HDDs over 5 years, with the most pessimistic assumptions possible.

- 80% Allocatable

OU Ceph Setup (cont)

- Similar setup planned for OU Health Sciences Center in Oklahoma City

- Also, planning CephFS Cache appliances

  ○ Cache Servers in various locations on campus

  ○ Will speed up read and write access to CephFS file system

XRootD Ceph Testing Status

- Created 10 TB Ceph partition, /xrd_test

- Mounted /xrd_test as CephFS file system on se1 (current SE)

- Brought up separate XRootD server for this CephFS partition

- Were able to access this CephFS partition just like XRootD storage via XRootD service

XRootD Ceph Testing

- Alternate XRootD server config very similar to prod config:

```
[hs@ouhep1 se]$ diff xrootd-cephfs.cfg xrootd-se.cfg
2d1
< xrd.port 64000
7,8c6,7
< all.export /xrd_test
< #pss.origin dms.oscer.ou.edu:1094
---
> all.export /xrd
> pss.origin dms.oscer.ou.edu:1094
11c10
< #ofs.osslib libXrdPss.so
---
> ofs.osslib libXrdPss.so
```

XRootD Ceph Testing continued

- Simple gfal testing:

```
gfal-copy --copy-mode pull https://se1.oscer.ou.edu:1094/xrd/srm-test/
twentyfivegiga https://se1.oscer.ou.edu:64000/xrd_test/atlasdatadisk/
twentyfivegiga1
```

- BNL FTS transfer tests:

```
Parallel transfer tests against
https://se1.oscer.ou.edu:64000/xrd_test/atlasdatadisk/
```

- Get up to 3.5 GB/s transfer speeds, but average closer to 1.5 GB/s

- Also, ATLAS production transfers going on at the same time

Current XRootD Ceph Testing

- Replicated XRootD configuration on new SE (currently se0.oscer.ou.edu)

- ceph file system not yet mounted on se0 because of issue new network routing

- So for now, test transfers go to local 960 GB SSD

```
gfal-copy -f --copy-mode pull https://se0.oscer.ou.edu:9090/ourdisk/hpc/
xrd_test/fivegiga https://se1.oscer.ou.edu:64000/ourdisk/hpc/xrd_test/
atlasdatadisk/fivegiga &
gfal-copy -f --copy-mode pull https://se1.oscer.ou.edu:64000/ourdisk/hpc/
xrd_test/atlasdatadisk/fivegiga https://se0.oscer.ou.edu:9090/ourdisk/hpc/
xrd_test/fivegiga &
```

- Don't have any actual network monitoring yet

- Ran 1 transfer of 5 GB file in about 7 s

- That amounts to around 700 MB/s transfer rate for single stream

- Ran 50 parallel transfers of 5 GB file in about 50 s

- That amounts to around 5 GB/s transfer rate

Summary and Conclusions

- Current (old) OU XRootD storage has been stable and performant for over 7 years

- Able to transfer 3+ GB/s, which is probably close to current available hardware/network limit

- Initial xrootd-ceph setup successful

- Looking forward to performance improvement with this new setup; hopefully get at least 50 Gbps transfer rates with new xrootd server

- Also still to do: Add Token Auth for XRootD