USATLAS-SWT2 Storage Evolution

Horst Severini

ACP 2023

September 2023

Outline

- Introduction
- Computing and Storage Hardware
- Network
- XRootD Configuration
- CephFS Testing and Migration Plans

Introduction

- USATLAS is the US Contingent of the ATLAS Collaboration at the Large Hadron Collider (LHC) at CERN
- There is one Tier-0 Center: CERN
- There are about a Dozen Tier-1 Centers around the World, and about 35 Tier-2 Centers
- The USATLAS Tier-1 Center is at Brookhaven National Lab (BNL)
- There are 4 USATLAS Tier-2 Centers: NET2, AGLT2, MWT2, and SWT2

Introduction

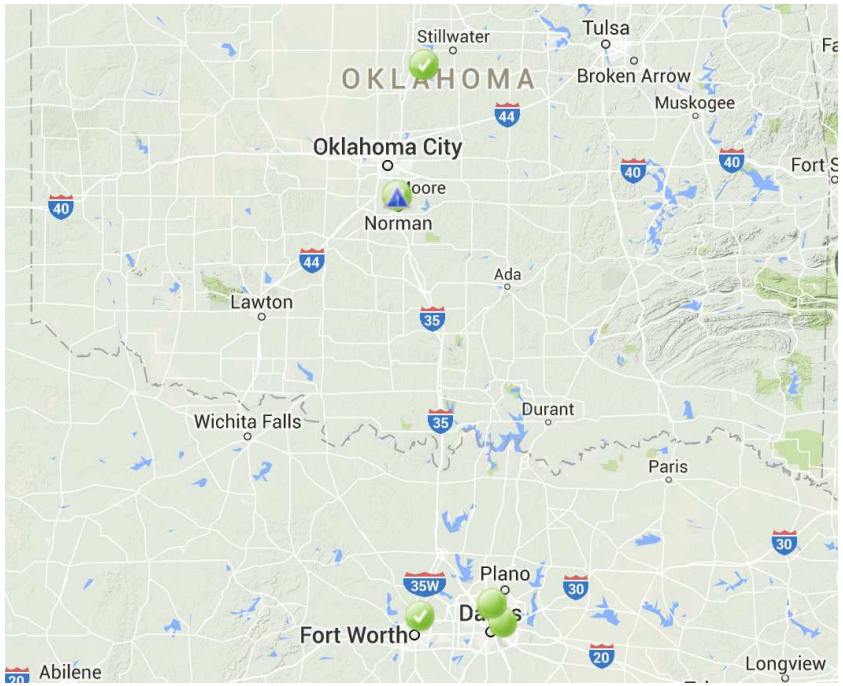
- University of Oklahoma (OU) is part of the USATLAS SWT2 Center, together with University of Texas at Arlington (UTA)
- Planning to migrate the OU storage from XRootD to Ceph in the next year
- XRootD and Ceph are both high performance File Systems
  - XRootD was developed specifically for High Energy Physics; has one central manager node and a number of storage nodes to distribute files
  - Ceph is a parallel file system which can deliver very high performance and redundancy/resiliency

File Systems Overview

- Linux/Unix stores files and directories (folders) on File Systems
- File Systems can be mounted Locally or Remotely
- Common Local File Systems are ext4, xfs, or zfs
- Standard Remote File System is NFS
- Throughput is limited by disk speed and network speed
- Parallel File Systems like XRootD and Ceph increase total Throughput with multiple disk servers

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 Grid GateKeeper, 1 DataTransferNode (StorageElement – 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 (CentOS 7.8), SLURM 22.5, OSG 3.6, XRootD 5.4.2
  - Upgrading the cluster to EL9 later this year
  - GateKeeper already running AlmaLinux 9.2

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)
  - ROCKS 7.0 (CentOS 7.6), SLURM 17.11.13, OSG 3.5

Network

- OU connected at 100 Gbps to Internet2 and ESnet via OneNet
- OSCER connected at 100 Gbps to OneNet
- Everything ipv6 ready except for HTCondor-CE GK
- SE on 100 Gbps DMZ – OFFN (Oklahoma Friction Free Network)
  - Dual 25-gig Bonded
  - iperf3 test over 40 Gbps
  - WAN XRootD transfers to storage nodes over 25 Gbps

XRootD Configuration

- Currently, se1.oscer.ou.edu acts as XRootD proxy 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
  - Not fully understood

Ceph Migration Plans

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

- 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

Current XRootD Ceph Testing Status

- Created 10 TB Ceph partition, /xrd_test
- Mounted /xrd_test as CephFS file system on se1 (current SE proxy)
- Brought up separate XRootD server for this CephFS partition
- Were able to access this CephFS partition just like XRootD storage via XRootD proxy service

XRootD Ceph Testing

- Alternate XRootD server config very similar to proxy 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
```

- Can get up to 400 MB/s transfer speeds
- 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
- Done without any tuning so far
- Also, ATLAS production transfers going on at the same time

Summary and Conclusions

- OU XRootD storage has been stable and performant for 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 closer to 50 Gbps current network limit
- Will get new high-end proxy server with 100 Gbps NIC for production
- Also still to do: Migrate from X509 to Token Auth for XRootD (Working fine on new EL9 HT-Condor-CE)