# OU XRootD Site Report

Horst Severini (University of Oklahoma)

XRootD and FTS Workshop 2026

# Introduction

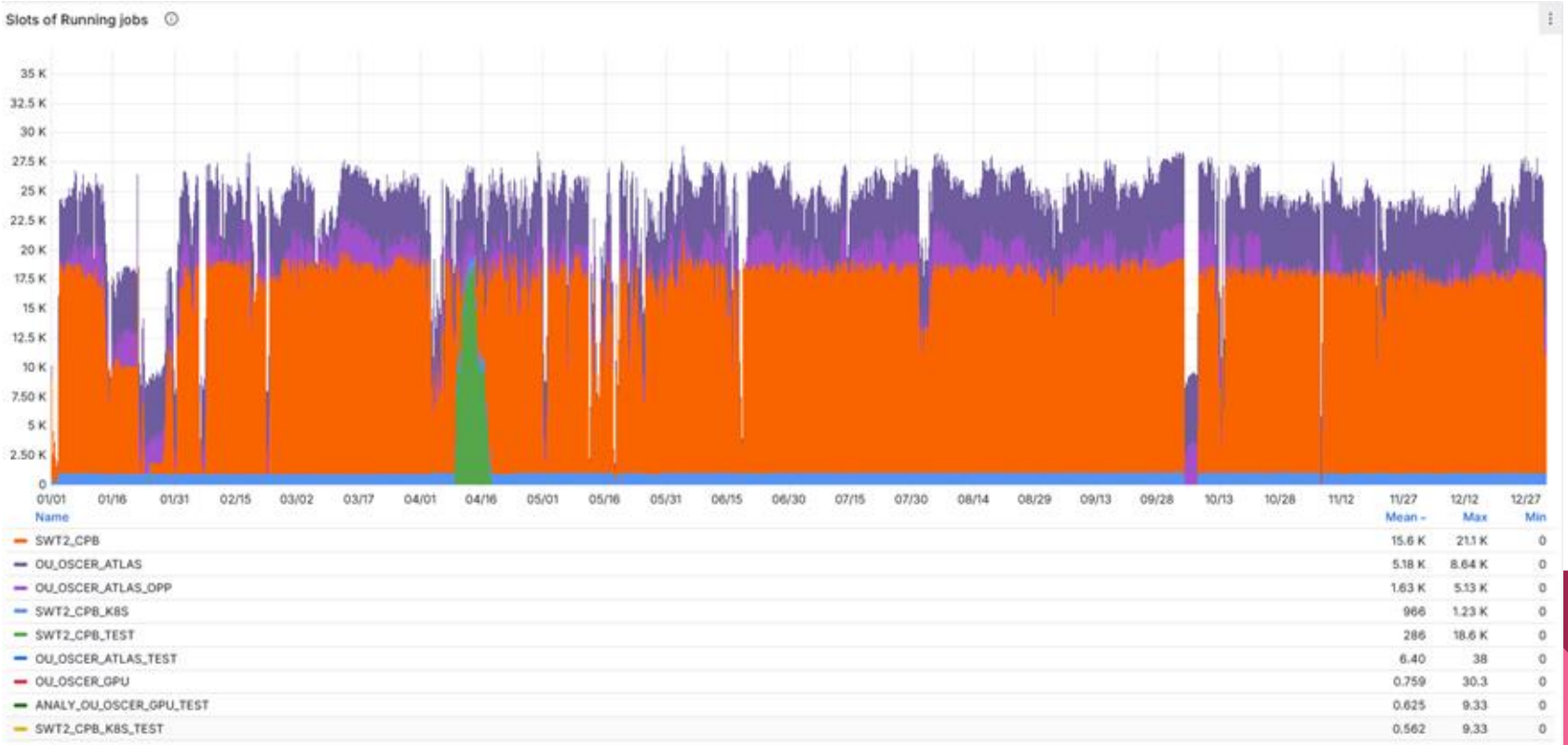
- The US ATLAS SouthWest Tier 2 (SWT2)
  - One of 4 ATLAS Tier 2's in the USA
  - Funded by The National Science Foundation (NSF)
  - Partnership between two universities in the "South West"
    - The University of Texas at Arlington (UTA)
    - The University of Oklahoma (OU)

U.S. National Science Foundation
Directorate for Mathematical and Physical Sciences

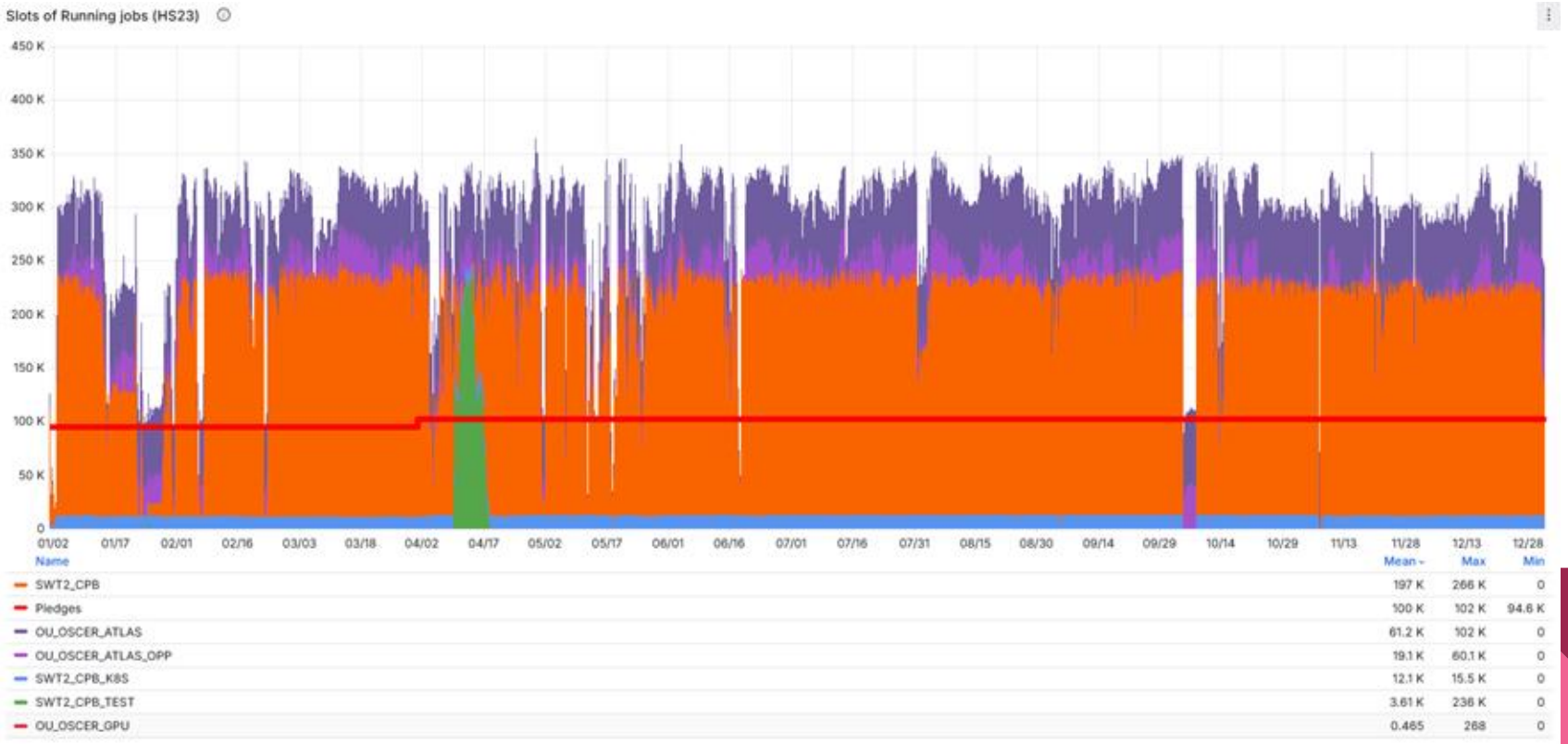
# Brief History

- Prototype UTA Tier 2 started with a NSF MRI grant awarded in 2002
- Participated in the first Data Challenge for ATLAS - DC0
- SWT2 selected as one of the first two National Cyberinfrastructure Tier 2 sites by US ATLAS around 2005 (NorthEast T2, PI: Shank, SWT2, PI: De)
- SWT2 was originally a partnership of UTA (lead), OU, University of New Mexico and Langston University - the last two are no longer participating
- At UTA, SWT2 moved from the University Data Center to a dedicated machine room in new CPB building around 2009
- At OU, part of OU Supercomputing Center for Education and Research (OSCER)

# SWT2 Usage in 2025 (slots)

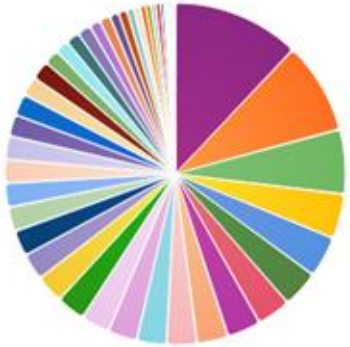


# For HS23 fans



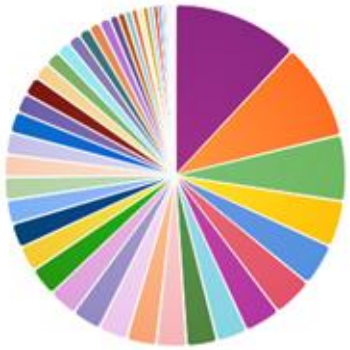
# Context within ATLAS (All Tier 1 and Tier 2 sites)

Slots of Running jobs (HS23)



| | Value | Percent |
|---|---|---|
| SI-SiGNET | 7.42 Bil | 12% |
| US-MWT2 | 5.33 Bil | 9% |
| US-T1-BNL | 3.73 Bil | 6% |
| US-SWT2 | 2.56 Bil | 4% |
| UK-T1-RAL | 2.37 Bil | 4% |
| CZ-Prague-T2 | 2.11 Bil | 4% |
| US-AGLT2 | 1.94 Bil | 3% |
| FR-CCIN2P3 | 1.93 Bil | 3% |
| DE-FREIBURGWUPPERTAL | 1.80 Bil | 3% |
| JP-Tokyo-ATLAS-T2 | 1.73 Bil | 3% |
| DE-KIT | 1.70 Bil | 3% |

Wall clock time. Successful jobs (HS23 seconds)



| | | |
|---|---|---|
| SI-SiGNET | 24.1 Tri | 12% |
| US-MWT2 | 18.4 Tri | 9% |
| US-T1-BNL | 12.6 Tri | 6% |
| US-SWT2 | 8.70 Tri | 4% |
| UK-T1-RAL | 7.97 Tri | 4% |
| US-AGLT2 | 7.63 Tri | 4% |
| FR-CCIN2P3 | 6.71 Tri | 3% |
| DE-KIT | 5.80 Tri | 3% |
| CZ-Prague-T2 | 5.69 Tri | 3% |
| JP-Tokyo-ATLAS-T2 | 5.64 Tri | 3% |
| DE-FREIBURGWUPPERTAL | 5.58 Tri | 3% |

# SWT2 Personnel

- UTA
  - Patrick McGuigan - retired in 2023
  - Mark Sosebee - retired in 2025
  - Zachary Booth - from industry - ramped up in 2024
  - Andrey Zarochentsev - from ALICE - joined 2025
- OU
  - Horst Severini
  - Chris Walker
- Other experts
  - Thanks to Armen Vartapetian for K8S, Fernando Barreiro for Google support
  - Very small but dedicated expert team

# PanDA Queues at SWT2

- SWT2_CPB (primary queue for both production and user analysis)
- OU_OSCER_ATLAS (primary queue for data production)
- OU_OSCER_ATLAS_OPP (opportunistic access to OSCER slots)
- OU_OSCER_GPU (opportunistic GPU queue with L40s, A100s)
- OU_OSCER_ATLAS_TEST (Testing new features like storage)
- Queues for various other testing purposes

# SWT2 Computing

- UTA
  - 459 compute nodes
  - 2 GKs - HTCondor-CE 24.0.1-1
  - SLURM 24.05.3-1
  - 2 Varnish (7.7.1-1) servers for Frontier (1 primary, 1 backup)
  - 2 Squid servers for CVMFS
- OU
  - 87 compute nodes (2 of them in OU_OSCER_ATLAS_TEST partition)
  - 1GK – HTCondor-CE 25.0.13-1
  - SLURM 23.02; about to upgrade to 26 for cgroups v2 support
  - 1 NRP onsite Varnish node for both CVMFS and Frontier

# Networking and Infrastructure

- OU Network upgrade
  - Currently 100 Gbps from OSCER to OU core and OFFN DMZ
  - From there to OneNet and GPN and Internet2 and ESNET and LHCONE
  - From OneNet out, dual 400 Gbps
  - Got CC* proposal for 400 Gbps OSCER core network funded
  - Now in the process of 400 Gbps upgrade to both OU Core and OFFN DMZ network

- UTA
  - Just upgraded local Core network from 30 to 80 Gbps

# SWT2 Storage

- We have ~17 PB available
  - 13.0 PB fully provisioned at UTA
  - 0.7 PB provisioned at OU
  - New 1 PB in UTA test cluster
  - New 1 PB at OU – migration in progress
  - Retired ~3.7 PB in past year (out of warranty)
  - Storage system is XRootD at both sites
    - UTA base system xfs
    - OU moving to CephFS for underlying storage, but staying with XRootD DTN

# Old XRootD Configuration

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

# OU Storage Upgrade/Migration

- Old XRootD storage to be retired
  - 7 100 TB T630s now out of warranty
  - Can get 2.5 GB/s throughput from WAN via XRootD DTN
  - Good for now, but needs to improve for long term

- New OURdisk CephFS storage
  - Purchased 1 PB of OURdisk storage, good for 7 years
  - Can get up to 3 GB/s throughput with fio from new single 100 Gpbs DTN to OURdisk partition
  - Can get up to 30 Gbps throughput from WAN via single XRootD DTN
  - Will add a second DTN to increase total WAN throughput into OURdisk to 50+ Gbps
  - Compute nodes have direct access to OURdisk like ‘local’ file system
  - Multi-stream iperf3 tests from BNL to OU close to 100 Gbps line speed

# OU Ceph Setup

- Migrate to CephFS file system now storage server warranty has expired
  - OU Research Disk – OURdisk
  - 21.5 PB and growing
  - In two separate Ceph clusters
  - 14 GB/s total throughput
  - Very performant and secure, reasonably priced
  - $93 per usable TB, good for 7 years

# OU Ceph Setup (cont)

- 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
- 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 Migration Status

- Have 1 PB Ceph partition, /ourdisk/hpc/xrd/
- Mounted as CephFS file systems on new DTN se0
- XRootD 5.9.6 server on new SE
- Accessing this CephFS partition like XRootD storage via XRootD service
  - Minimal changes to XRootD configuration
- Modified configuration to run as usatlas1 rather than xrootd user
  - That way, files on CephFS system automatically correct owner
- Can be directly accessed from compute nodes
- Issues
  - Still having SELinux issues with running xrootd server as usatlas1 rather than xrootd user

# Transfer Testing

- Using iperf3
  - Can get close to 100 Gbps line speed between BNL and OU
  - Needed to bring up 5 iperf3 servers in parallel for that
  - So, WAN network is fine
- Using dd
  - Ran dd transfer tests from se0 and se2 to both cephfs clusters
  - Can get between 2 and 3 GB/s throughput for each dd transfer
  - Can get up to 5 GB/s combined throughput, either se0+se2 to either cluster, or one se to both clusters
  - Will use se0+se2 for production, and 1 PB partition on one cephfs cluster
- Using XRootD
  - Hiro got 30 Gbps xrootd transfers from BNL to OU, using se0
  - Production Rucio DDM transfers got up to 4 GB/s throughput
  - Now in the process of moving from old to new storage
  - Still some issues with local ‘copytool’ for stage-in/out

# Rucio DDM Migration

- Working with Rucio team on DDM Migration
  - Changed copy-tool from xrdcp to cp
  - Still some issue with local ‘copytool’ for stage-in/out
  - Adjusted CRIC settings
  - WAN inbound and outbound transfers are fine
  - Most data has been copied over from old to new storage by Rucio

# Conclusions

- XRootD Migration at OU a long time coming
- Almost done now
- Should be able to get 5-6 GB/s throughput between two DTNs
- Working with Rucio team to finalize DDM Migration
- Will add additional DTNs later as needed to keep up with transfer needs
- Add Round Robin DNS for SEs