MWT2 Site Report

Judith Stephen on behalf of the MWT2 team
Enrico Fermi Institute
University of Chicago

MANIAC LAB

US ATLAS Facilities Meeting at SLAC
November 30, 2022

MWT2

- Site overview and operations
- Procurement plans and retirements
- Milestones
- Contributions
- Concerns/issues

MWT2 Personnel

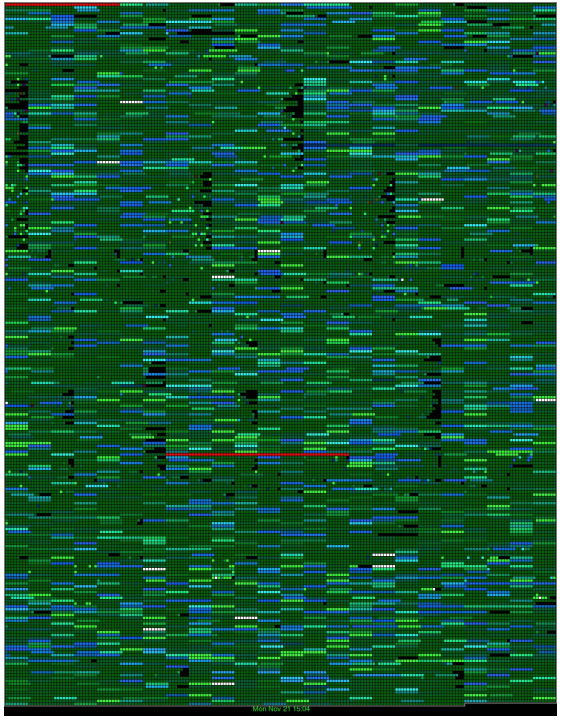
- PIs: Rob Gardner, Fred Luehring, Mark Neubauer
- Staff: Ed Dambik, Farnaz Golnaraghi, Fengping Hu, David Jordan, Judith Stephen





MWT2 Site Overview

- Compute
  - HTCondor-CE 5.1.5, HTCondor 9.0.13
  - 41k cores across three sites
- Storage
  - dCache 7.2.15
  - 47 pool nodes, 1 head node, 3 VMs for doors
  - 15.466 PB
- Analytics (see David's talk)
  - Elasticsearch 8.3.2
  - 6 head nodes, 17 data nodes, 1 kibana node
  - 132 TB total, 46 TB used. All SSDs
- Shared Tier 3 (see Fengping's talk)

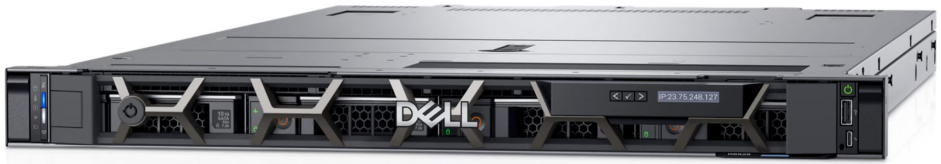


Live view of MWT2 CPU utilization

MWT2 Server Ops in a Nutshell

1. Operating System installation via **Cobbler**

2. Configuration via **Puppet** + **Git**



3. Monitoring via **Ganglia** and **Graphite**

4. Alerting via **Nagios** and **Slack** integration

5. Remote management via **IPMI**

Other Tools/Services Supporting MWT2

- KVM and OpenStack
  - All virtualized services
- SLATE/Kubernetes nodes for edge services
  - IU + UC
  - Federated services: XCache, Squid
- Six perfSONAR systems
- ZFS-based NFS /home and backup services
- NetBox for inventory management
- SysView for HTCondor pool monitoring

Security for MWT2 and Tier3/AF

- Crowdstrike and Rapid7
  - Using University site-wide licenses
  - Crowdstrike Falcon – anti-malware and intrusion detection
  - Rapid7 InsightVM – vulnerability management
- MFA for the Analysis Facility, Google Authenticator
  - Plan to roll out MFA for Tier3/AF users in 2023
- Remote Access
  - SSH Hardening (AllowedGroups, Key-based Auth, ProxyJump)
  - Investigating Yubikeys (hardware tokens) for admins
- Monitoring, management behind UChicago firewall
  - Per recommendations from UChicago IT Services, services that aren't heavy data movers go behind Campus border firewall rather than SciDMZ

Site-wide Linux Tunings

- Ulimit changes to allow jobs and storage to access large numbers of file descriptors
- Storage tunings such as read-ahead block sizes and queue depths
- ESNet FasterData tunings applied to all nodes, including also increasing NIC TX queue lengths and other optimizations
- Custom kernels on workers especially those with BCM57414 and Intel XK710 NICs as we have noticed many subtle network-related job failures that seem to be remedied by newer kernels
    - Probably will be removed in EL9

Procurement

- MWT2 procurement plan link
- Retirements
  - 7.7 PB of dCache storage is out of warranty
    - 2016, 2018
  - Upcoming UIUC retirements
    - 34 nodes / 1520 cores to be retired, no remaining extensions
    - 40 nodes / 2336 cores out of warranty but can be extended

Storage Pilot at IU

- 6x retired R720s + MD1200 shelves from UC
  - Initial deployment of 1 R720 + 6 shelves
  - Keep at least 1 R720 and shelf+disks as spares
  - 3TB disks currently. Could retrofit with larger disks
- Considering different options
  - Add to dCache (locality cache?)
  - IU XCache

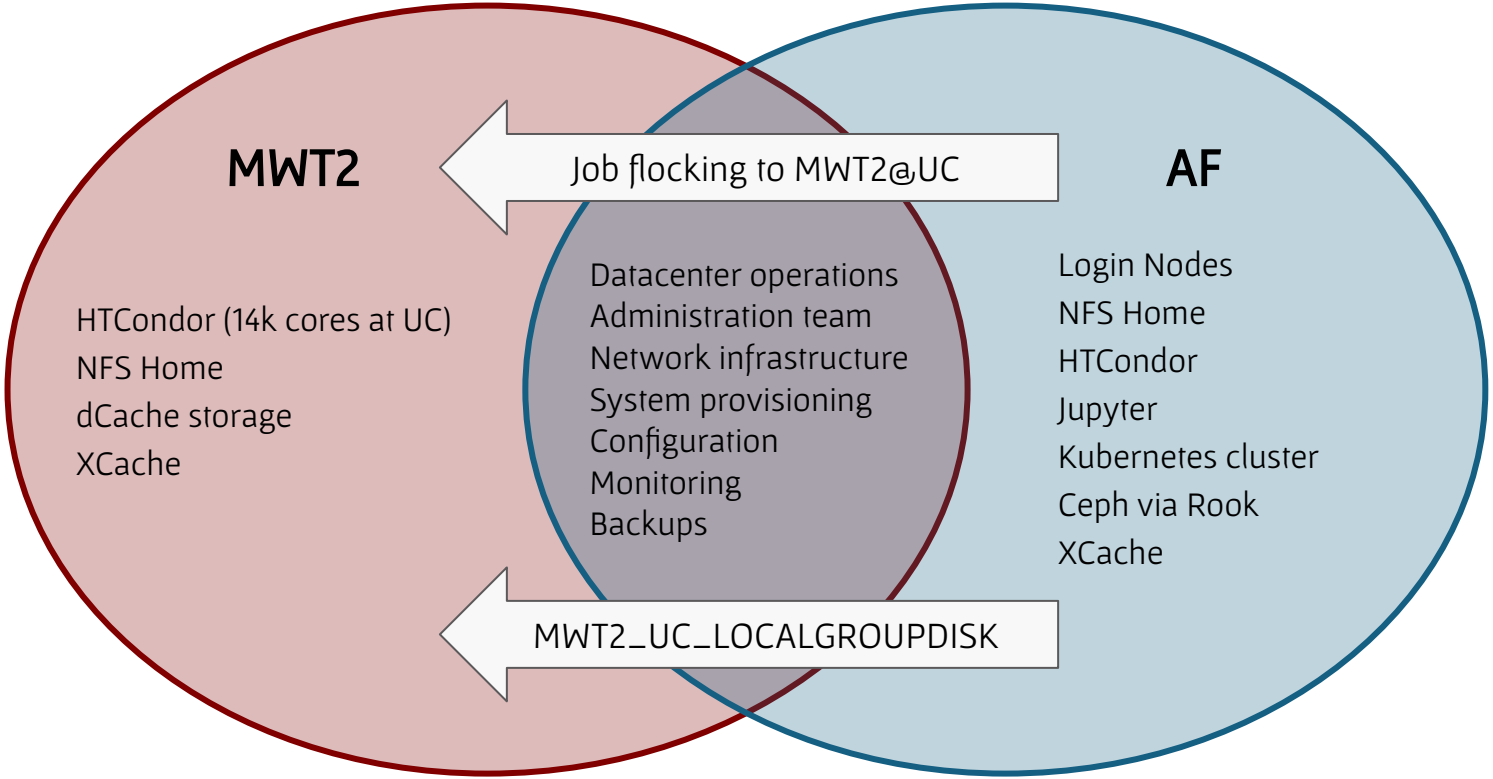
Milestones

- OS decisions
  - Will migrate to some flavor of EL9, skipping EL8
- Software updates
  - Apptainer roll-out completed
  - HTCondor 10 in progress, including configuration refactor, TOKEN-based pool authentication, and accounting groups reorganized
  - HTCondor-CE 6 when available
  - dCache 8.2 FY23 Q1
- WLCG Security Operations Center implementation at MWT2 (Zeek)
  - Planned for FY23 Q1/Q2

Site Contributions to US ATLAS

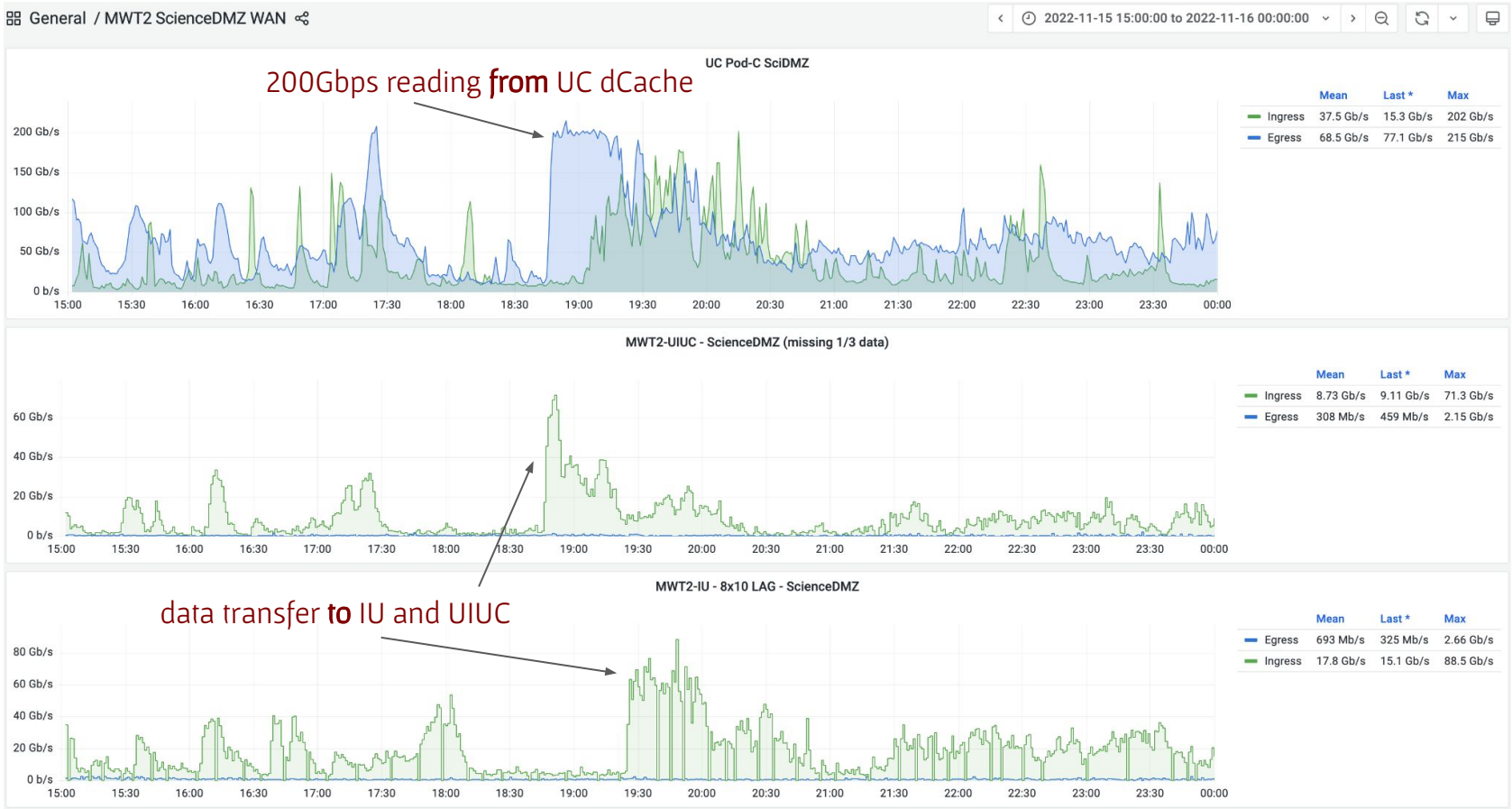
- Shared Tier3/Analysis Facility
  - Shared infrastructure, administrative support, user support, training events, Analysis Grand Challenge, ServiceX and analysis facility validation
    - See Fengping's talk
- Analytics Platform, metrics
  - See Ilija's and David's talks
- Cloud operations, troubleshooting (Fred)
- Federated Ops for XCache and Frontier Squid
- Work with OSG-LHC
  - Software stack testing
- Facility R&D with IRIS-HEP SSL, FABRIC/FAB
- Container image registry (Harbor), security and distribution – SOTERIA

Analysis Facility Integration



Site Concerns and Issues

- Network bandwidth limitations from UC to IU and UIUC when recovering from HC resets/drains and downtimes
- Occasional site drainage/offlining reasons not always clear
- While MWT2 has not had security problems, we want to proactively improve our security posture
- Storage redundancy, particularly for storage that is out of warranty


General / MWT2 ScienceDMZ WAN
2022-11-15 15:00:00 to 2022-11-16 00:00:00
UC Pod-C SciDMZ
200Gbps reading from UC dCache
Mean Last * Max
Ingress 37.5 Gb/s 15.3 Gb/s 202 Gb/s
Egress 68.5 Gb/s 77.1 Gb/s 215 Gb/s
200 Gb/s
150 Gb/s
100 Gb/s
50 Gb/s
0 b/s
15:00 15:30 16:00 16:30 17:00 17:30 18:00 18:30 19:00 19:30 20:00 20:30 21:00 21:30 22:00 22:30 23:00 23:30 00:00
MWT2-UIUC - ScienceDMZ (missing 1/3 data)
Mean Last * Max
Ingress 8.73 Gb/s 9.11 Gb/s 71.3 Gb/s
Egress 308 Mb/s 459 Mb/s 2.15 Gb/s
60 Gb/s
40 Gb/s
20 Gb/s
0 b/s
15:00 15:30 16:00 16:30 17:00 17:30 18:00 18:30 19:00 19:30 20:00 20:30 21:00 21:30 22:00 22:30 23:00 23:30 00:00
MWT2-IU - 8x10 LAG - ScienceDMZ
data transfer to IU and UIUC
Mean Last * Max
Egress 693 Mb/s 325 Mb/s 2.66 Gb/s
Ingress 17.8 Gb/s 15.1 Gb/s 88.5 Gb/s
80 Gb/s
60 Gb/s
40 Gb/s
20 Gb/s
0 b/s
15:00 15:30 16:00 16:30 17:00 17:30 18:00 18:30 19:00 19:30 20:00 20:30 21:00 21:30 22:00 22:30 23:00 23:30 00:00

Troubleshooting drains… (Fred's analysis)