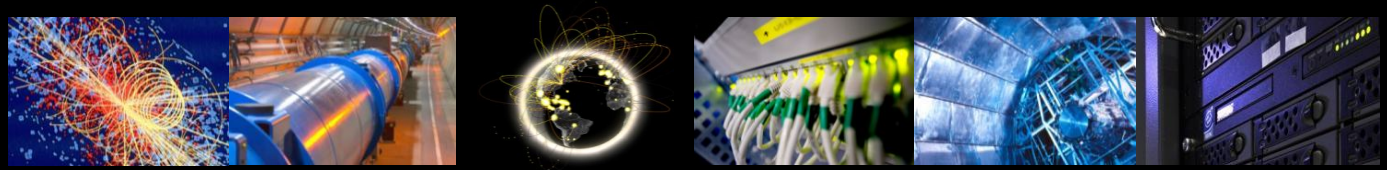
Results from Fall 2024 US Mini-Challenges

Diego Davila / University of California San Diego & Shawn McKee / University of Michigan

WLCG DOMA General Meeting (https://indico.cern.ch/event/1495675/)

January 15, 2025



WLCG Worldwide LHC Computing Grid

Open Science Grid

Fall 2024 US Mini Challenges

As previewed in the Nov 13, 2024 WLCG DOMA, both USATLAS and USCMS undertook some capacity mini-challenges, designed to benchmark our current infrastructure.

These were simple load-tests where we wanted to evaluate the capacity limits for our various sites.

We were not trying to identify where we might adversely interact with other activities, as we do when we run the regular data challenges.

The fall challenges were orchestrated by Hiro Ito / BNL (for USATLAS) and Diego Davila / UCSD (for USCMS).

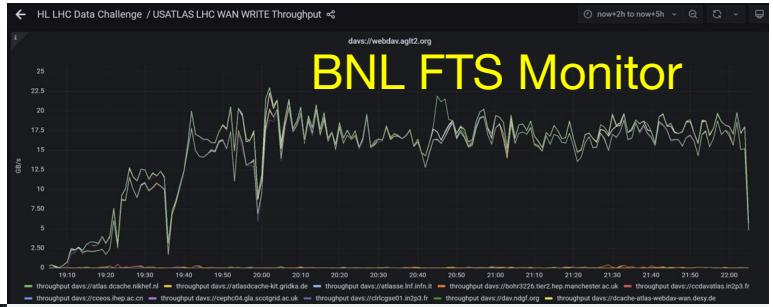
Original plan: USATLAS Mini Data Challenge Fall 2024

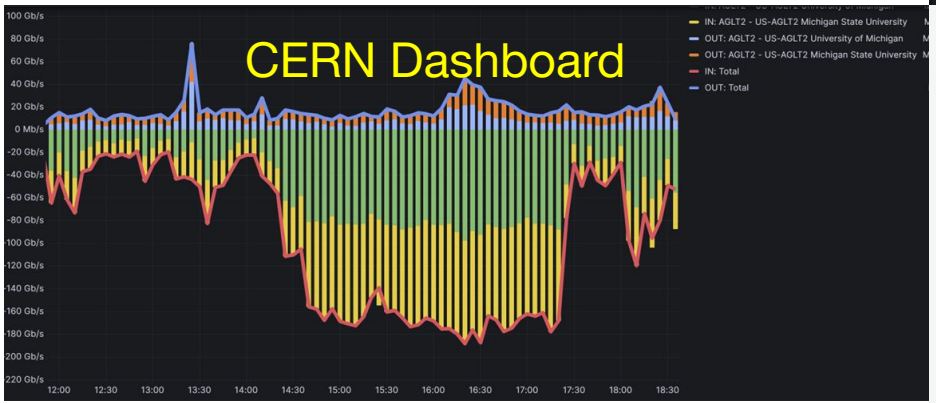
- Use load test[1] from T1 to each T2s at full T2's network capacity
  - To check if there are any changes from the results from the last test.
  - Network capabilities of US T2s: AGLT2(200 Gbps) MWT2 (200 Gbps), NET2 (expected to be 400 Gbps). SWT2 (100 Gbps)
  - Individually as well as simultaneously
    - Simultaneous test might present "choke" point in the path.
- T2s to T1 at full wan disk capacity.
  - Not capable to reach the full network capability of BNL at 1.6 Tbps due to the storage layout of T1 storage
- T1 Tape staging and readout test.
  - Check the staging throughput and readout throughput of staged data from BNL.
- Check and validate the accuracy of the various monitor at the site as well as the central ones at CERN, ESNet, BNL,...

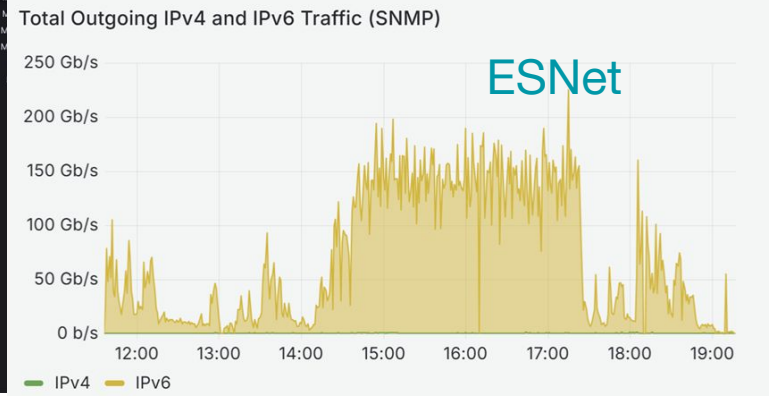
[1] The program is found at the following BNLBox folder https://bnlbox.sdcc.bnl.gov/index.php/s/XGs6LJEGNzf69zK

AGLT2 Ingestion Testing

- The observed throughput for injecting AGLT2 was about 150Gbps.
- Various monitors were checked against each other to evaluate their accuracies.
- Although all monitor shows the similar number, CERN Dashboard seems a bit higher the other two? However we must note that CERN Site Network is ALL traffic

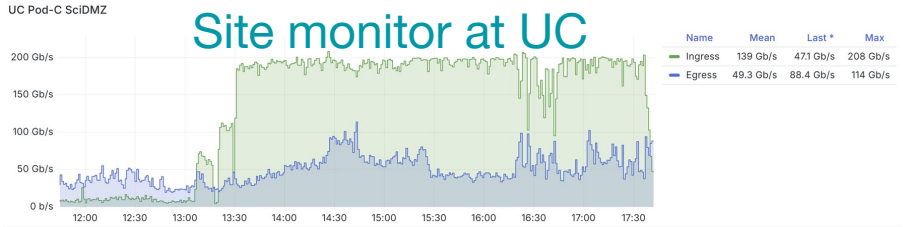


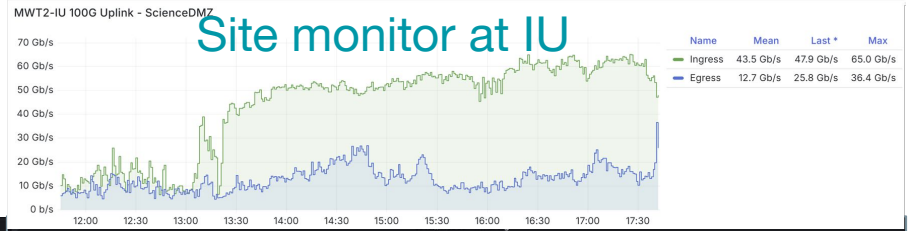


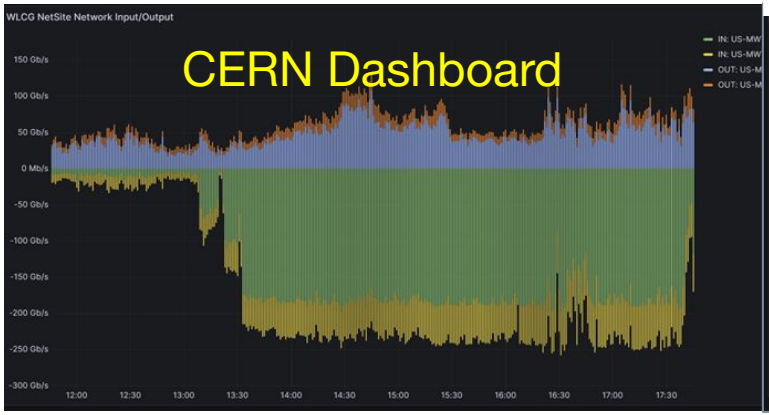


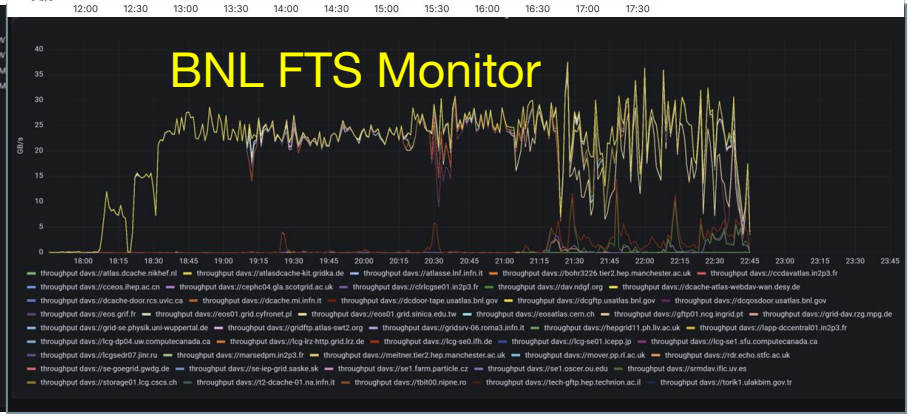
MWT2 Ingestion Testing

- The observed throughput for MWT2 was about 200Gbps.
- CERN Dashboard shows again a bit higher values.
- NOTE: ESNet monitor only shows UC.



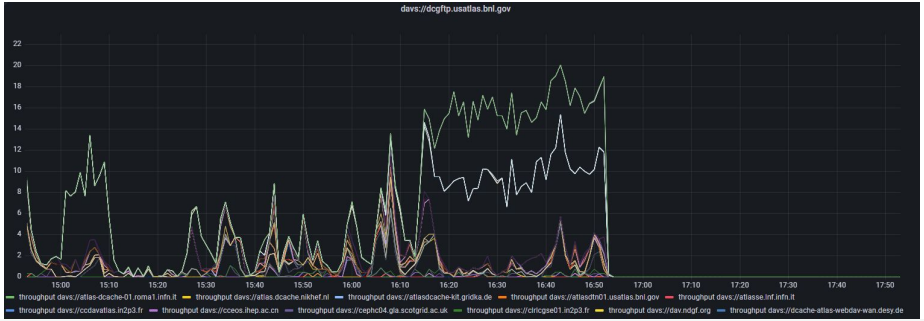


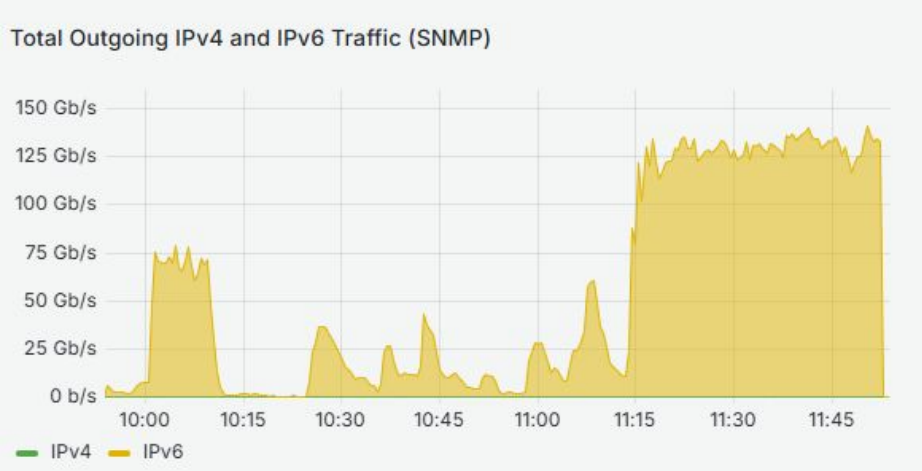




BNL T1 Ingestion Testing

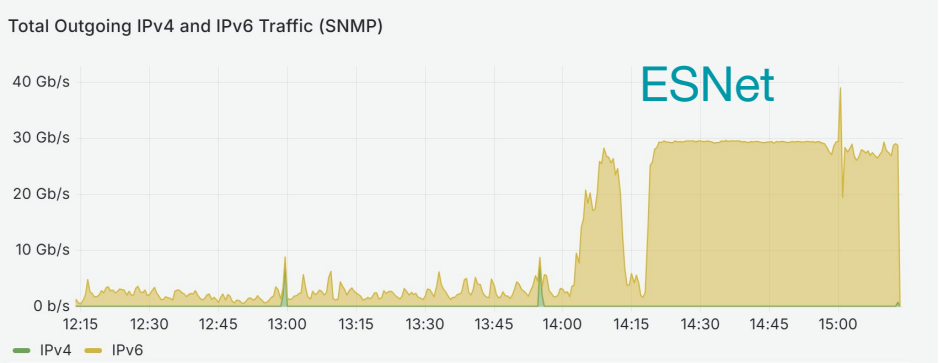
- Complication
  - It requires multiple sites to drive BNL to its bandwidth capacity
  - AGLT2 encountered storage issue at the time of BNL testing.
    - Cause of delay
  - Some shorter testing after AGLT2 became operational.
- It achieved ~125Gbps.
- It requires additional testing to investigate the actual current limitation. (Redo in February?)





SWT2

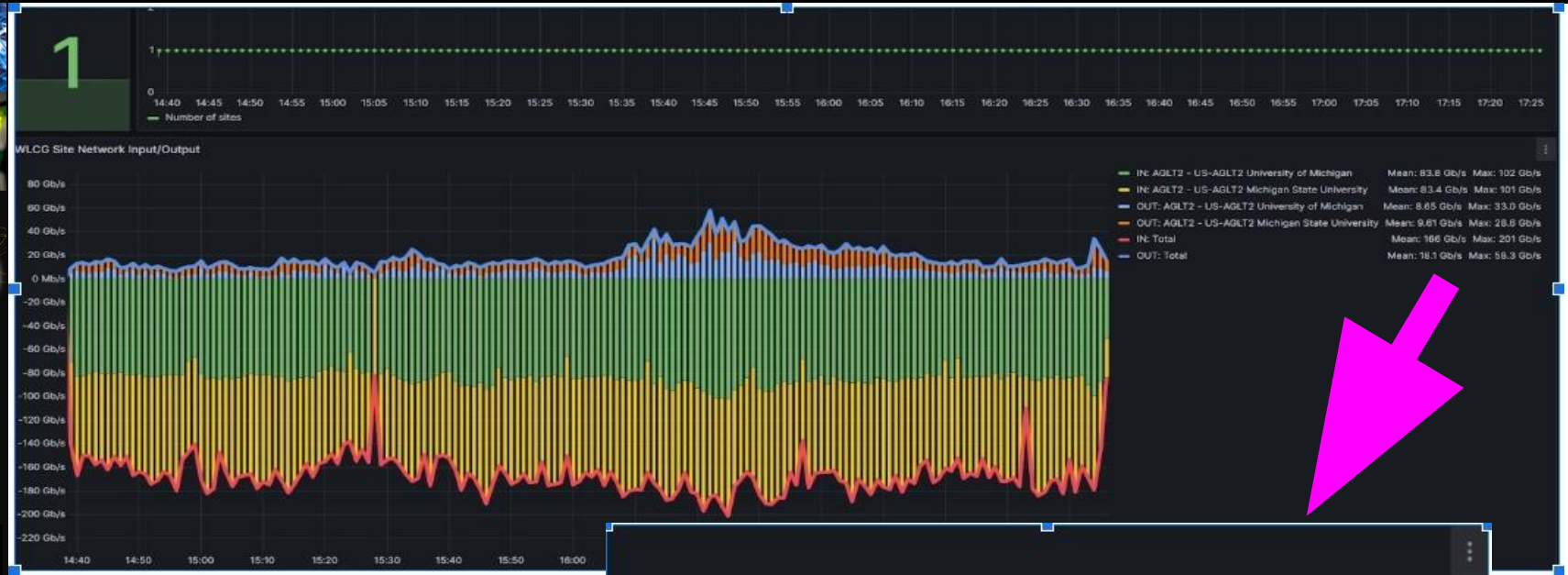
- SWT2 (UTA) has achieved 30 Gbps.
- This is still the limit of the network at the site.
  - The flatness of the plot indicates that it is indeed the network limit.
- Discussion with the network engineer is under way to increase the bandwidth. (Needs both bandwidth and DTN capacity increases)



USATLAS Testing Summary

- Summary of injection testing to US ATLAS sites
  - AGLT2 has achieved 150Gbps.
  - MWT2 has achieved 200Gbps.
  - BNL has achieved 125Gbps.
    - In addition to the disk throughput, the analysis of tape system in terms of the staging throughput is on going.
      - Staging test will be conducted soon (in Jan/Feb).
  - SWT2 UTA has achieved 30Gbps.
  - NET2 was not quite ready for the testing.
    - It is waiting for the completion of the network upgrade to 400Gbps
    - It will be tested as soon as the site is ready (next few weeks?)

AGLT2 Site Report: Mini DC traffic in 4 hour window



IN: AGLT2 - US-AGLT2 University of Michigan Mean: 83.8 Gb/s Max: 102 Gb/s
IN: AGLT2 - US-AGLT2 Michigan State University Mean: 83.4 Gb/s Max: 101 Gb/s
OUT: AGLT2 - US-AGLT2 University of Michigan Mean: 8.65 Gb/s Max: 33.0 Gb/s
OUT: AGLT2 - US-AGLT2 Michigan State University Mean: 9.61 Gb/s Max: 28.6 Gb/s
IN: Total Mean: 166 Gb/s Max: 201 Gb/s
OUT: Total Mean: 18.1 Gb/s Max: 58.3 Gb/s

AGLT2 Site Report

- The UM network (80 Gbps) is saturated, but MSU (100 Gbps) it not (84% used)
- **UM site**: No bottlenecks observed from border switches, rack switches, dcache head nodes (postgresql), dCache head/pool nodes (cpu load, memory usage, disk IO performance)
- UM site: 33 pool nodes (R740xD2 and R760xD2) with 66 pools , storage nodes monitor shows DC traffic added an avg of 300MB/s to each node (300MB/s*33*8=79.2Gbps, so the limit comes from the 80 Gbps link)
  - There is still plenty of room for more IO based on historical peak IO. We need **~264 Gbps WAN** in order to saturate the storage nodes' IO capacity.
- **MSU site**: has similar hardware and quantity implying we should be able to source/sink **~500 Gbps WAN** if we had that link capacity available.
- Detailed site report available

Diego will now cover the USCMS results

USCMS Load Test Tool Overview

We used the same tool used in DC24: dc_inject:
https://gitlab.cern.ch/wlcg-doma/dc_inject

It receives 3 main inputs:

1. A **list of unique datasets** in each of the Sites used as Sources. This is calculated using a separate script prior to the tests.
2. A Json file with the Source/Dest pair of sites to be tested and the **desired rate**
3. The **injection period** (default: 15min)

The tool picks from the **list of datasets** enough data to achieve the **desired rate** for the **desired period** and use them to create short-lived rules in Rucio

USCMS DC26 mini-challenges Plans for Fall (1 of 3)

Two Main Goals:

1. Get all sites to report to the WLCG monitoring dashboard:
https://monit-grafana-open.cern.ch/d/MwuxgogIk/wlcg-site-network?orgId=16&from=1730827738666&to=1731432538666

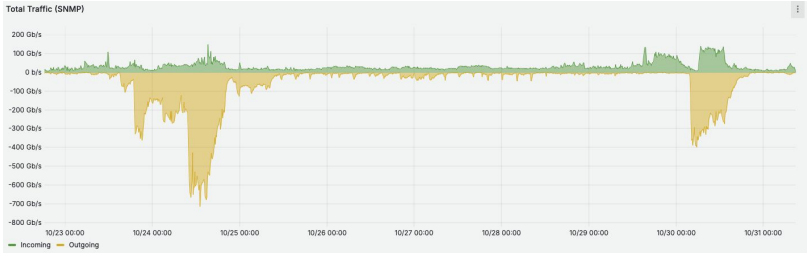
2. Load Test all T2s and FNAL at the highest rate proposed for DC24:

- T2: ~100 Gbps
- FNAL: ~400 Gbps
- We can increase if Sites are ready to push harder

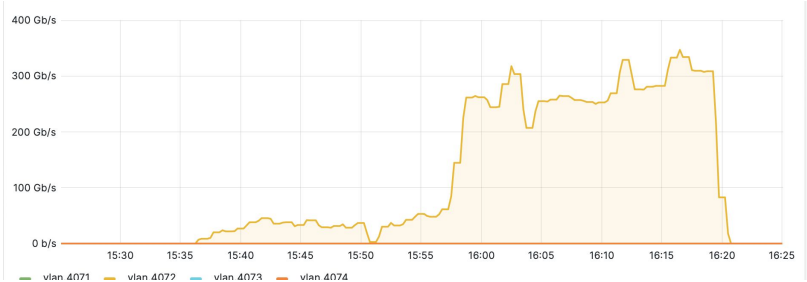
USCMS DC26 mini-challenges Plans for Fall (3 of 3)

Skipped Sites

FNAL was already Load Tested by Production at ~700Gbps:
https://dashboard.stardust.es.net/d/xAueBcH7k/lhc-data-challenge-interface-details?orgId=2&var-iface=fnalfcc-cr6%3A%3Afnal_se-1600&from=1729646326003&to=1730389887908



Caltech has already demonstrated ~300 Gbps:
https://indico.cern.ch/event/1343110/contributions/6065564/attachments/2939546/5163932/go



UCSD has an ongoing network upgrade :(

USCMS DC26 mini-challenges Plans for Fall (3 of 3)

Load Test Plan:

- Select 2 consecutive days for each T2
- Perform a round of Load Tests between a given T2 and FNAL
- Analyze and Adjust if necessary
- Perform another round the following day
- Collect plots

USCMS Testing Summary

		OUT (Site → FNAL)		IN (FNAL → Site)	
Site	**Max Target**	**Max Achieved**	**Best day Average**	**Max Achieved**	**Best day Average**
Nebraska	125	100	80	100	90
Wisconsin	150	160	100	75	60
Florida	300	190	100	200	135
Vanderbilt	100	100	90	100	80
MIT	80	70	65	30	28
Purdue	100	100	100	70	60

Please takes these numbers with a grain of salt.

Max Achieved: more or less stable Maximum

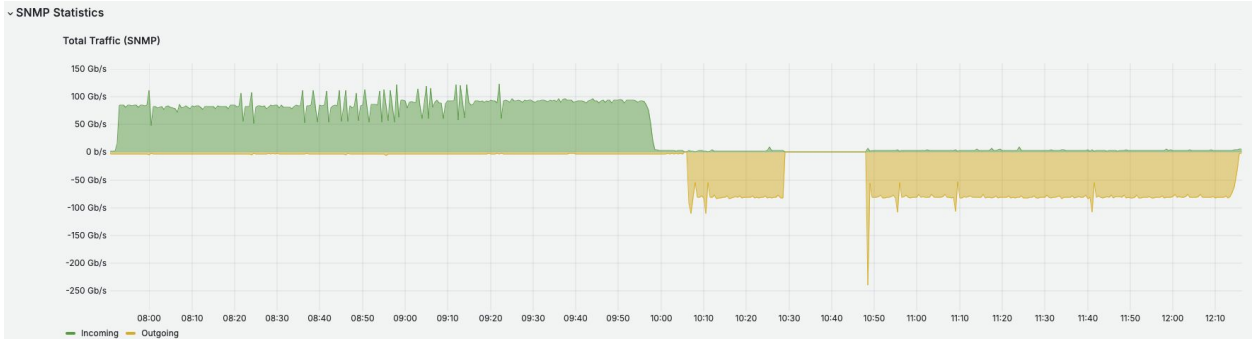
Best day Average: more or less sustained Average

Dealing with throughput is tricky (see next slide)

Full report available here:
https://docs.google.com/document/d/1rtfhxfsvHYfqc5xQXsFQot-Vu_Ek83WGJMEfKCiVs1I/edit?usp=sharing

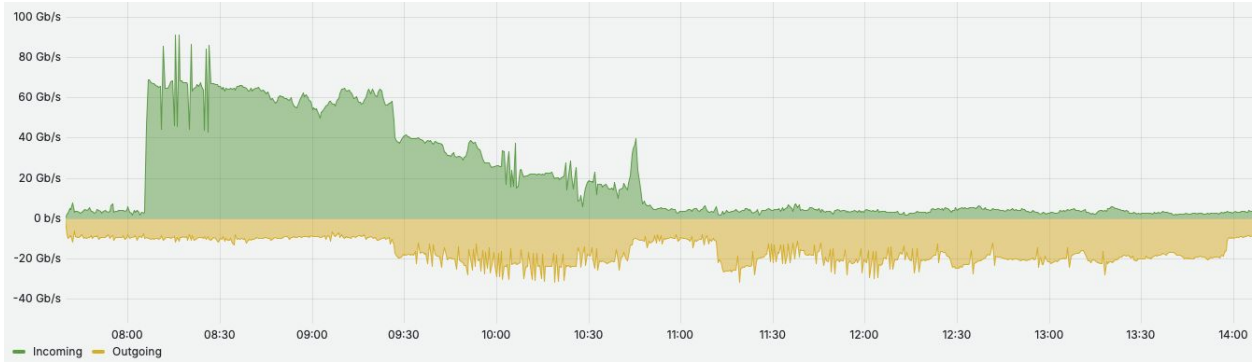
USCMS: Dealing with throughput is tricky

Some plots are easy to analyze



.. others not so much.

What should we consider green's average ?



USCMS Testing Highlights (1 of 2)

		OUT (Site → FNAL)		IN (FNAL → Site)	
Site	**Max Target**	**Max Achieved**	**Best day Average**	**Max Achieved**	**Best day Average**
Nebraska	125	100	80	100	90
Wisconsin	150	160	100	75	60
Florida	300	190	100	200	135
Vanderbilt	100	100	90	100	80
MIT	80	70	65	30	28
Purdue	100	100	100	70	60

big diff between read/write rates

big diff between max and avg. pushing harder didn't help

Upcoming upgrade so not too worried about low results

Reads are great, writes not so much

USCMS Testing Highlights (2 of 2)

		OUT (Site → FNAL)		IN (FNAL → Site)	
Site	**Max Target**	**Max Achieved**	**Best day Average**	**Max Achieved**	**Best day Average**
Nebraska	125	100	80	100	90
Wisconsin	150	160	100	75	60
Florida	300	190	100	200	135
Vanderbilt	100	100	90	100	80
MIT	80	70	65	30	28
Purdue	100	100	100	70	60

Almost all sites can sustain reads at ~100 Gbps

Significant variation in write rates

USCMS WLCG Monitoring status

Site	Status	Observations
Caltech	OK	
Florida	OK	Fixed by Swapping IN and OUT in the SNMP script
MIT	Missing	They haven't contacted their Network team to request SNMP access
Nebraska	OK	
Purdue	OK	
UCSD	OK	Fixed by filtering unwanted connections. Https version cannot handle more than 1 connection at a time
Vanderbilt	OK	Deployed the go version. They had issues with how the output was being handled but this has been fixed by monIT as of this week.
Wisconsin	OK	
FNAL	OK	Reconfigured after upgrading their Border Router

USCMS 2024 Fall mini-challenge Outcomes

- Found a bottleneck at Nebraska (issue is understood)
- Found a problem with the go implementation of the WLCG Monitoring
- Found missing checksums at Florida
- Increased FTS limits to 300+ from the 200 default for most T2s
- Switched: Florida and Wisconsin to use FNAL's FTS instead of CERN's
- Found an asymmetry within ESnet for Purdue: reads/writes over different interfaces.

Summary & Plans

We have successfully tested current USATLAS and USCMS capacity during a set of Fall 2024 site measurements, identifying some issues for further work.

We (IRIS-HEP/OSG-LHC/US-LHC) need to further **clarify** and **document** existing plans, **mini-challenges** and goals for the next year and onwards to DC26/DC27 (Capabilities next up in February!)

We have an **opportunity** to leverage DC24 and following results to improve our infrastructure, to drive technology deployment, to show value and to demonstrate capabilities at scale.

Questions or Discussion?

Acknowledgements

Thanks to **Hiro Ito** for his contributions to the slides and for running the USATLAS tests!! and to **Asif Shah** for helping running the USCMS tests.

We would like to thank the **WLCG**, **HEPiX**, **perfSONAR** and **OSG** organizations for their work on the topics presented.

In addition we want to explicitly acknowledge the support of the **National Science Foundation** which supported this work via:

- IRIS-HEP: NSF OAC-1836650 and PHY-2323298

Background Material

Here are some resources we know about:

Presentations

- WLCG Data Challenge 2024 (DC24) Status and Plans Related to ATLAS DDM (Jun 2023)
- DC24 Planning and Near Term Activities (Jul 2023)
- USATLAS Data Challenge 2024 Take-aways (Feb 2024)
- Medium to Long Term Network Plans for ATLAS and CMS (Mar 2024)
- DC24 Network Activities & Results (May 2024)

Some Google Docs

- WLCG/DOMA Data Challenge 2024: Final Report
- USATLAS Milestones/MiniChallenges for Next WLCG Data Challenge in 2024
- Planning Mini-Challenges for US ATLAS Facilities and Distributed Computing
- NOTES: USATLAS Facility Status and Evolution Discussion

DC24 Links

Official DC24 report
https://zenodo.org/records/11402618

DC24 Network Activities and Results:
https://docs.google.com/presentation/d/1s0VvbXEpj1PN9umFT8wgsHsHmG9EYucymbalKNrvuKQ/edit#slide=id.p1

Katy Ellis LHCONE/LHCOPN DC24 presentation:
https://docs.google.com/presentation/d/1Tm3pCMkfHj5KHTW3PXbgS7mdHf72Ir27qr1JgMbrnRg/edit#slide=id.g1ea89411ecb_0_4

Next Steps Towards DC26:
https://docs.google.com/presentation/d/1mMx6QaihWJWpbVEQgxNjZXRT5_s4SkBTXu0SpELtuvI/edit#slide=id.gd170caf633_1_0

DC24 ATLAS Retrospective:
https://docs.google.com/presentation/d/1Lh_D57BvWn13AFClhhuczm-j-tKV-yMez_oD4yYUtBo/edit#slide=id.gd170caf633_1_0

Backup Slides

OSG-LHC/IRIS-HEP Current Plans

At the IRIS-HEP retreat in September 2024, we discussed how to prepare for DC26
As mentioned, mini-challenges are an important tool that we want to enable
Goals for the next DC:

- Move the majority of our data via IPv6 and have one or more sites **IPv6-only**
- Have 80%+ of our traffic identified by SciTags
- Have SENSE/Rucio used in production at one or more sites
- Improve site network monitoring to identify traffic by LHCONE, LHCOPN, R&E and commodity

The plan:

- (DONE) Before the end of 2024 rerun capacity tests for US sites to determine current values
- (NEXT) February 2025, execute a joint USATLAS-USCMS **capabilities** mini-challenge: scitokens, SciTags, SENSE, jumbo frames
- Early-to-mid Summer 2025, execute a joint USATLAS-USCMS **capacity** mini-challenge