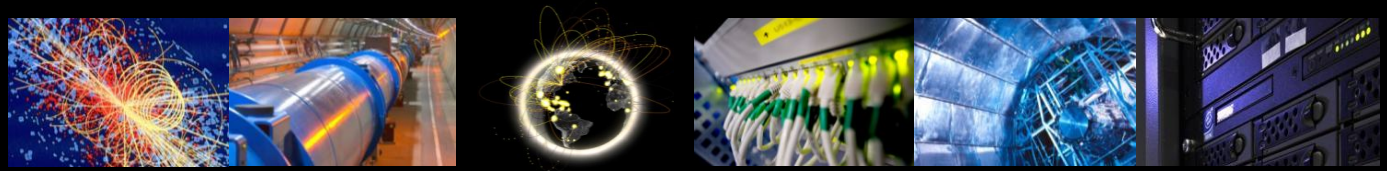
Results of Mini Data Challenges in the U.S.

Garhan Attebury / UNL, Diego Davila / UCSD, Hiro Ito / BNL, Shawn McKee / UMich

WLCG OTF #6 (https://indico.cern.ch/event/1562124/)

September 10, 2025



WLCG Worldwide LHC Computing Grid

Open Science Grid

Overview

The IRIS-HEP project has been informing US plans and activities regarding WLCG Data Challenges and associated “mini-challenges”. We plan for two general types of challenges:

Capacity mini-challenges: Capacity (throughput) of data transfers to benchmark our sites and identify bottlenecks.

Capability mini-challenges: Testing application and technology capabilities we foresee as helping us to improve our infrastructure and meet HL-LHC goals.

In this presentation we want to report on the August US capacity mini-challenge…

Plan History

The IRIS-HEP retreat plan agreed upon in fall 2024 (in conjunction with USATLAS/USCMS) was the following:

- Fall 2024, run targeted capacity tests to benchmark our sites
- January 2025, run a regional capacity test to identify areas of collision
- Summer 2025, run targeted capability tests

By November 2024 we modified the plan, keeping the ongoing capacity tests but changing the order of the other two items

- February 2025, target various capability tests, proponent organized and run
- Early-to-mid summer 2025, run larger scale capacity tests, adding any new beneficial capabilities identified in February (this became the August test)

August US Capacity Mini-Challenge

We created a Google doc to discuss, plan and execute the capacity test:

https://docs.google.com/document/d/1RiTDBMR2xRnjLa2tGT_kvGLfTaDBUfHPUXpnoPftnjc/edit?tab=t.0

Many details are recorded there for reference.

Overall idea was to use the last two week's in August 2025 for testing. The first week would allow us to "benchmark" each site individually while the second would focus on joint testing to identify any bottlenecks.

Note all testing done while running normal production workloads

USATLAS August Testing Plan Overview

We had two goals:

1. run a new “load-test” measurement to determine the current bottleneck rate for each site (IN/OUT)
2. Identify any network bottlenecks shared with other experiments

We planned to use Hiro’s custom load-testing scripts to move synthetic data between sets of sites (uses FTS but not Rucio).

Sites were asked to monitor performance and look for bottlenecks.

Joint testing period for the second week was planned in conjunction with USCMS.

USATLAS Schedule: First Week

We allocated August 18-22 to do individual site testing for ATLAS and CMS

USATLAS Testing

Eastern Time	Mon 18	Tue 19	Wed 20	Thur 21	Fri 22
09-11		BNL->AGLT2			
11-13		AGLT2->BNL	BNL->UTA UTA->BNL		
13-15		BNL->MWT2	BNL->NET2		BNL->OU OU->BNL
15-17		MWT2->BNL	NET2->BNL		

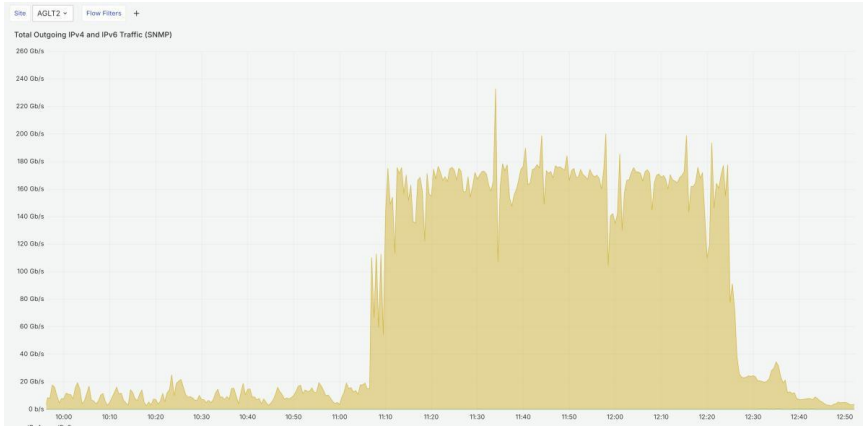
We utilized the Tier-1 to run tests to/from each Tier-2

From BNL to AGLT2

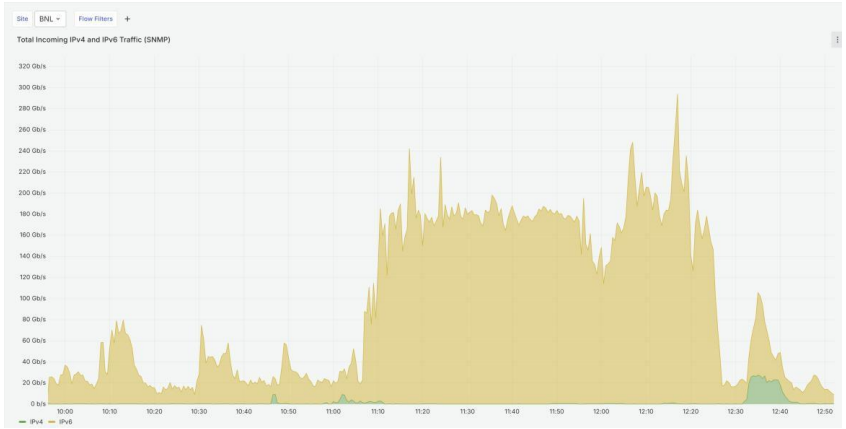
Test: From BNL to AGLT2
Physical limit: AGLT2 is limited at 180 Gbps

Results: BNL can push data to AGLT2 at AGLT2's site capacity limit of 180Gbps

AGLT2 (in)



BNL (out)



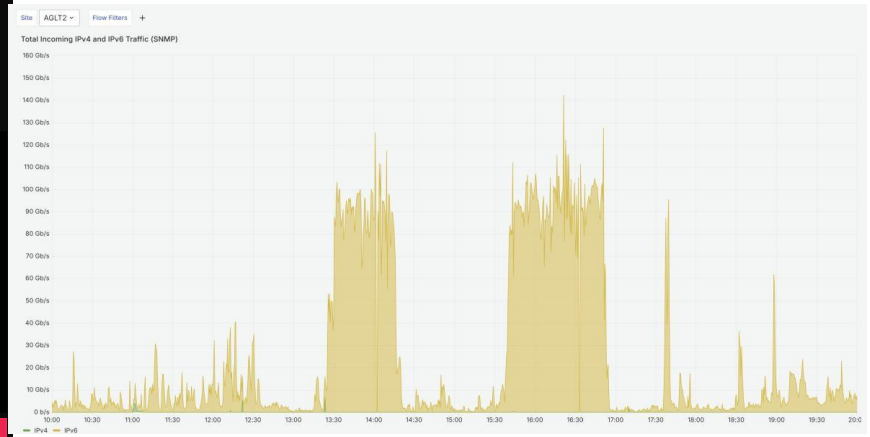
NOTE: Graph level is reverse. The plots do not show the data rate corresponds exclusively between AGLT2 and BNL. However, since the graph shows the good match, the big fraction of the data corresponds to the data between BNL and AGLT2.

From AGLT2 to BNL

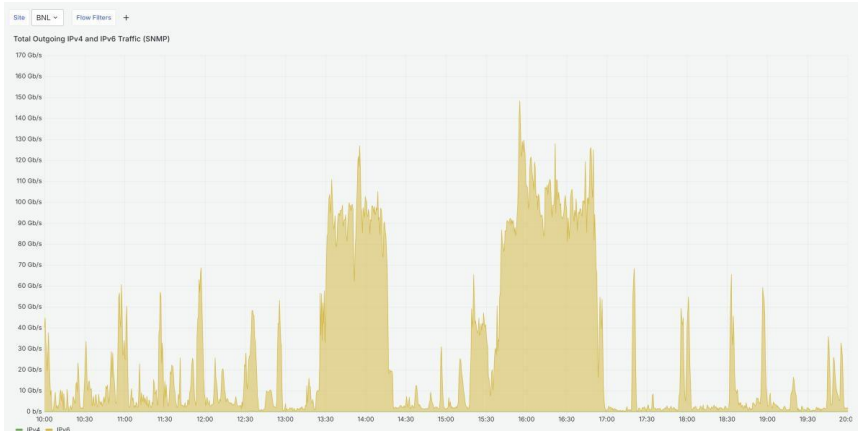
Test:	From AGLT2 to BNL
Limit:	AGLT2 is limited at 180 Gbps

Results: at the current configuration of BNL dCache using the intermediate pools on the door, 100 - 150 Gbps is the injection rate at BNL with reasonable number of concurrency (800).
NOTE: It is possible to push the rate further by increasing concurrency to very large number (~ 3000) . However, it has the diminishing return on per-file-rate

AGLT2 (out)



BNL (in)



NOTE: Graph level is reverse. The plots do not show the data rate corresponds exclusively between AGLT2 and BNL. However, since the graph shows the good match, the big fraction of the data corresponds to the data between BNL and AGLT2.

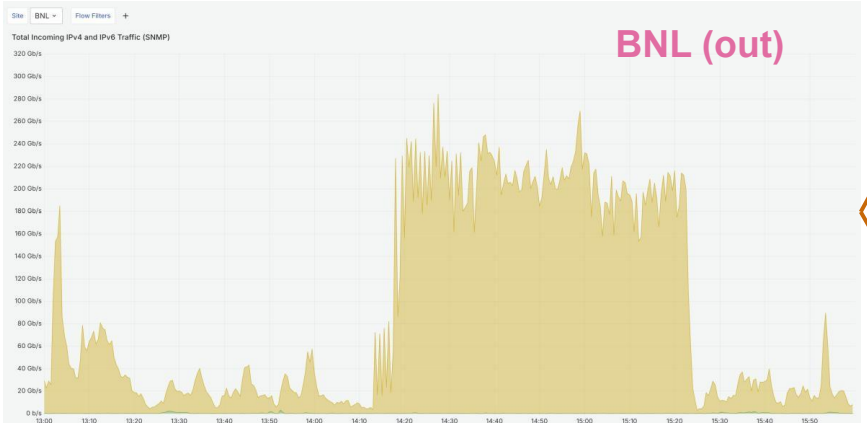
From BNL to MWT2

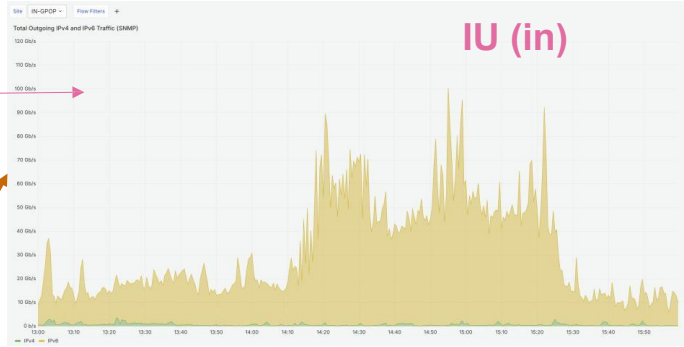
Test: From BNL to MWT2
Physical limit: MWT2 is limited at 300 Gbps

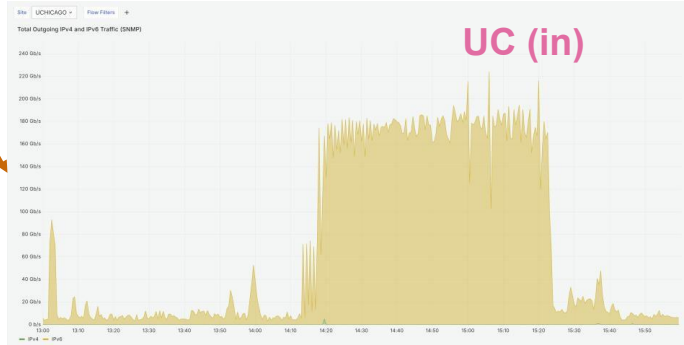
Results: BNL can push data to MWT2 at MWT2's site to about 250Gbps.

It doesn't saturate IU.

Pool selection in dcache might be preventing.







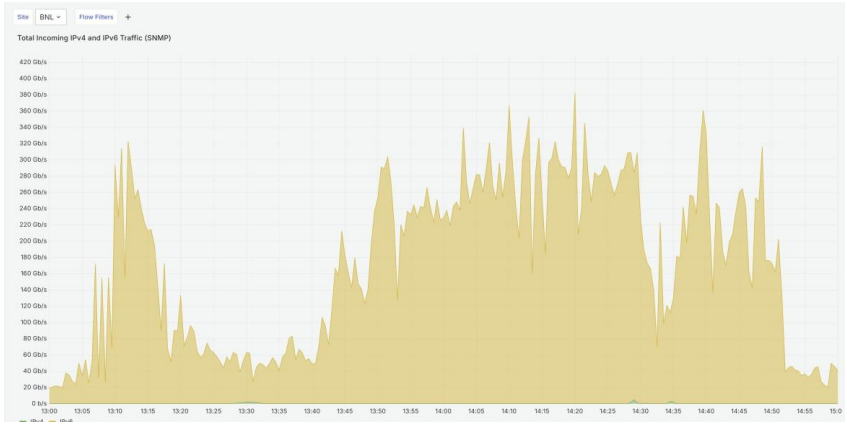
From BNL to NET2

Test: From BNL to NET2
Physical limit: NET2 is limited at 400 Gbps

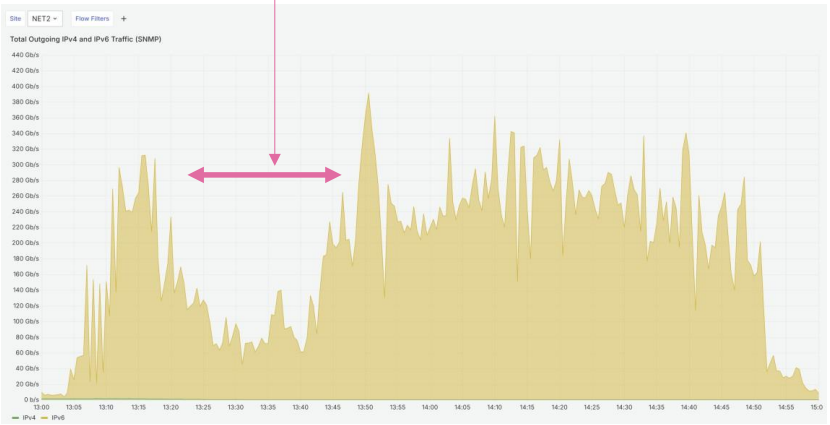
Results: BNL can push data to NET2 at NET2's site to about 380Gbps.

It doesn't saturate constantly.
Pool selection might be issue.
Number of concurrency was increased during the test after seeing the dip in the transfer rate.

BNL (out)



NET2 (in)

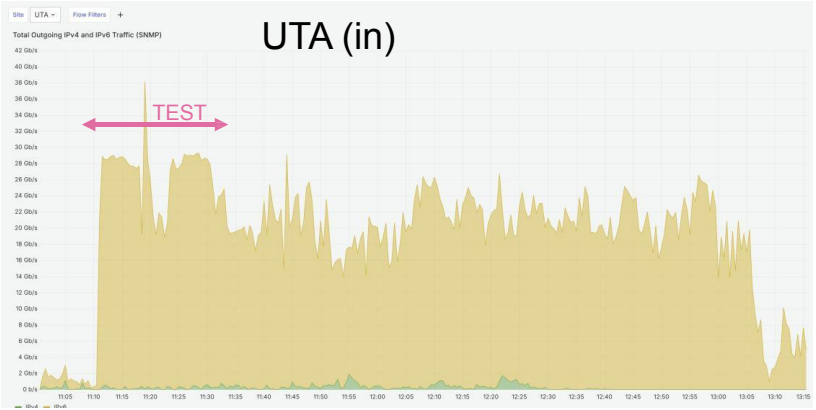


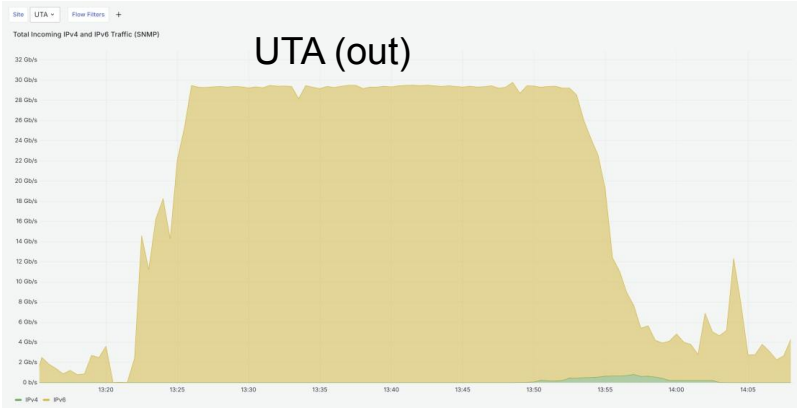
From and To SWT2 (UTA)

Test: To and from SWT2 (UTA)
Physical limit: SWT2 (UTA) is limited at 30 Gbps

Results: SWT2(UTA) is able to obtain their site maximum WAN rate for both directions.

NOTE: The duration of the injection rate is kept short since the production is already using the large fraction of the available bandwidth.



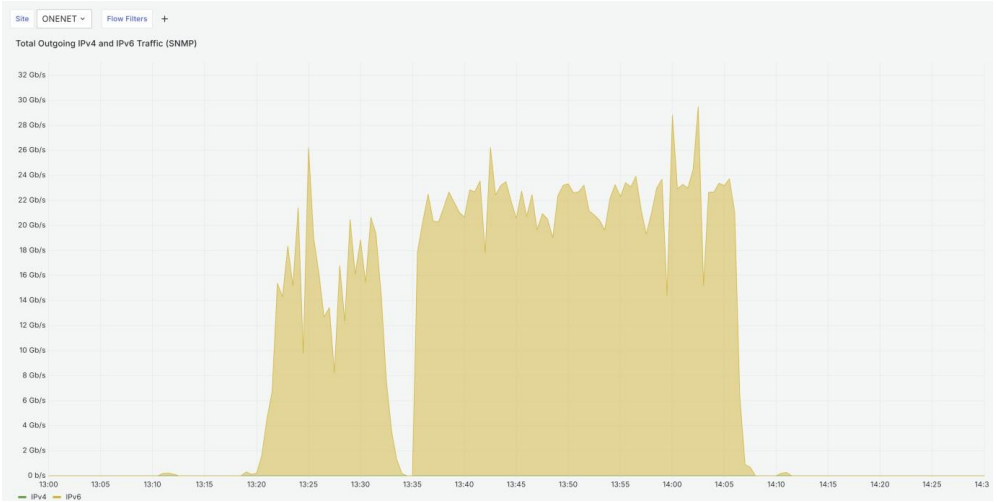


To SWT2 (OU)

Test: To and from SWT2 (OU)
Physical limit: SWT2 (OU) is limited at 30 Gbps

Results: SWT2(OU) is able to sustain close to the maximum WAN rate for injection.

NOTE: Because the new storage hardware and the network equipment is arriving to the site, only injection test is conducted. Also, the test didn't push the rate too hard due to the use of the older hardware.



USATLAS Comparing Fall 2024 with Summer 2025

Comparing: Previous vs Current capacity mini-challenge

Site	Fall 2024	Summer 2025	2025/2024
AGLT2	150 (180)	180 (180)	**120%**
MWT2	200 (200)	250 (300)	**125%**
NET2	NP (10)	380 (400)	**-**
SWT2	30 (60)	55 (60)	**183%**
BNL	200 (1600)	200 (1600)	**100%**

Numbers in parenthesis are the physical wirespeed possible for each site
Numbers in the table are basically “writes” to each site since BNL still passes data via dCache doors
SWT2 results have OU in the Summer 2025 testing but not in the Fall 2024 testing
NET2 wasn’t able to participate in the Fall test.

Summary: General improvement observed

USCMS August Testing Plan Overview

Similarly to USATLAS, we had 2 main objectives for this mini-challenge:

1. Test the individual capacity of each USCMS Site
2. Look for possible places of collision with USATLAS

The T2 target rates were selected by each of the Site admins to be somewhere between 100 and 200 Gbps. For FNAL we calculated a rate of 800 Gbps based on the DC original numbers.

For the individual Phase we used the *dc inject* tool to create ephemeral Rucio rules at the desired rate in one direction e.g. FNAL => Nebraska, then waited for the test to wind down and later do the opposite direction (Nebraska => FNAL).

As usual, we used FNAL to test each individual T2 and all the T2s to test FNAL

USCMS Schedule: First Week

Central Time	Mon 18	Tue 19	Wed 20	Thur 21	Fri 22
08-10	FNAL=> Nebraska	Florida=> FNAL	Caltech=> FNAL	FNAL Reads	Vanderbilt=> FNAL
10-12	Nebraska =>FNAL	FNAL=> Florida	FNAL=> Caltech	FNAL Reads	FNAL=> Vanderbilt
12-14	Purdue=> FNAL	Wisconsin=> FNAL	MIT=> FNAL	FNAL Writes	UCSD=> FNAL
14-16	FNAL=> Purdue	FNAL=> Wisconsin	FNAL=> MIT	FNAL Writes	FNAL=> UCSD

Some of the tests above had to be repeated the following week due to diverse circumstances. All results presented here are the best attempt of each site.

USCMS Week 1 Results (1/6)

		Reads		Writes	
Site	**Target**	**Max**	**Avg**	**Max**	**Avg**
FNAL	800	350	270	818	636
Caltech	100	90	85	96	70
Florida	200	52	50	45	35
MIT	100	85	80	62	47
Nebraska	200	109	65	145	115
Purdue	150	189*	180*	120	70
UCSD	100	47	45	48	50
Vanderbilt	200	170	120	60	50
Wisconsin	200	120	65	86	83

(*) During the Read test at Purdue, data was pulled from Purdue's XCache into UNL to put extra pressure on their Network

Max is the maximum value observed either in the WLGC or the Stardust dashboard.

Avg is the maximum value sustained for at least 10 minutes

USCMS Week 1 Results (2/6)

Comparison of Target vs Average Rates

Site	Target	Reads (Avg)	Reads/Target	Writes (Avg)	Writes/Target
FNAL	800	270	**34%**	636	**80%**
Caltech	100	85	**85%**	70	**70%**
Florida	200	50	**25%**	35	**18%**
MIT	100	80	**80%**	47	**47%**
Nebraska	200	65	**33%**	115	**58%**
Purdue	150	180	**120%**	70	**47%**
UCSD	100	45	**45%**	50	**50%**
Vanderbilt	200	120	**60%**	50	**25%**
Wisconsin	200	65	**33%**	83	**42%**

Comparing the expected rate (Target) with the Avg results we can conclude that either:

- We were too optimistic OR
- Many sites are really broken or falling behind

USCMS Week 1 Results (3/6)

BUT.. if we compare this mini-challenge with the previous one (8 months ago) things look a bit better:

- the majority of the test improved
- some site are obviously broken (should be easy to fix)

	Writes Avg			Reads Avg		
Site	**Fall 2024**	**Summer 2025**	**2025/2024**	**Fall 2024**	**Summer 2025**	**2025/2024**
Florida	135	35	**26%**	100	50	**50%**
MIT	28	47	**168%**	65	80	**123%**
Nebraska	90	115	**128%**	80	65	**81%**
Purdue	60	70	**117%**	100	180	**180%**
Vanderbilt	80	50	**63%**	90	120	**133%**
Wisconsin	60	83	**138%**	100	65	**65%**

N.B. Many sites did not participate in the previous mini-challenge

USCMS Week 1 Results (4/6)

Comparing: Previous vs Current mini-challenge

	Writes Avg			Reads Avg		
Site	**Fall 2024**	**Summer 2025**	**2025/2024**	**Fall 2024**	**Summer 2025**	**2025/2024**
Florida	135	35	**26%**	100	50	**50%**
MIT	28	47	**168%**	65	80	**123%**
Nebraska	90	115	**128%**	80	65	**81%**
Purdue	60	70	**117%**	100	180	**180%**
Vanderbilt	80	50	**63%**	90	120	**133%**
Wisconsin	60	83	**138%**	100	65	**65%**

N.B. Many sites did not participate in the previous mini-challenge

So bad that something must be broken

USCMS Week 1 Results (5/6)

Comparing: Previous vs Current mini-challenge

	Writes Avg			Reads Avg		
Site	**Fall 2024**	**Summer 2025**	**2025/2024**	**Fall 2024**	**Summer 2025**	**2025/2024**
Florida	135	35	**26%**	100	50	**50%**
MIT	28	47	**168%**	65	80	**123%**
Nebraska	90	115	**128%**	80	65	**81%**
Purdue	60	70	**117%**	100	180	**180%**
Vanderbilt	80	50	**63%**	90	120	**133%**
Wisconsin	60	83	**138%**	100	65	**65%**

N.B. Many sites did not participate in the previous mini-challenge

We have a good idea of what happened here

USCMS Week 1 Results (6/6)

Comparing: Previous vs Current mini-challenge

	Writes Avg			Reads Avg		
Site	**Fall 2024**	**Summer 2025**	**2025/2024**	**Fall 2024**	**Summer 2025**	**2025/2024**
Florida	135	35	**26%**	100	50	**50%**
MIT	28	47	**168%**	65	80	**123%**
Nebraska	90	115	**128%**	80	65	**81%**
Purdue	60	70	**117%**	100	180	**180%**
Vanderbilt	80	50	**63%**	90	120	**133%**
Wisconsin	60	83	**138%**	100	65	**65%**

N.B. Many sites did not participate in the previous mini-challenge

No idea of what happened here...

yet

USCMS Week 1: Story Time (1/6)

Florida

- During the test their Network Engineers were engaged and reported:
  - "All seems good, traffic is in the right route"
- From ESnet we got that: they are transitioning from one ESnet interface (Nash) to a new one (Wash)
  - "Would be good to test again after"
- Weird metrics seen in their perfsonar e.g. negative OWDs or 3ms to UCSD

Source	Destination	IP Version	Packet Loss (Max)	Throughput (Min)	Latency (Min)
perfsonar2.rc.ufl.edu	psonar9-191-2.fnal.g...	6	2.00%		-13.0 ms
perfsonar2.rc.ufl.edu	psonar-1.t2.ucsd.edu	6	0.167%		3.44 ms
perfsonar2.rc.ufl.edu	psl02-gva.cern.ch	6	0%		21.9 ms
perfsonar2.rc.ufl.edu	ps-lat.cms.rcac.purd...	4	0%		-18.5 ms
perfsonar2.rc.ufl.edu	perfsonar01.hep.wis...	6	0.333%		-2.64 ms
perfsonar2.rc.ufl.edu	perfsonar01.cmsaf....	6	25.3%		-15.3 ms
perfsonar2.rc.ufl.edu	perfsonar-lt.sprace....	6	0%		-17.5 ms

USCMS Week 1: Story Time (2/6)

Nebraska

- SENSE issue
  - They recently enabled multi-subnet support for SENSE.
  - For some yet-to-be-understood reason XRootD started using all the available subnets.
  - The new subnets were not registered in CRIC, and thus not picked up by FNAL as LHCONE ranges
  - 3 quarters of the traffic ended up going outside the LHCONE

- Discards
  - They noticed a significant O(1k/s) discard rate
  - Adding more capacity between datacenter core switches and DTN ToR switch fixed the issue.

USCMS Week 1: Story Time (3/6)

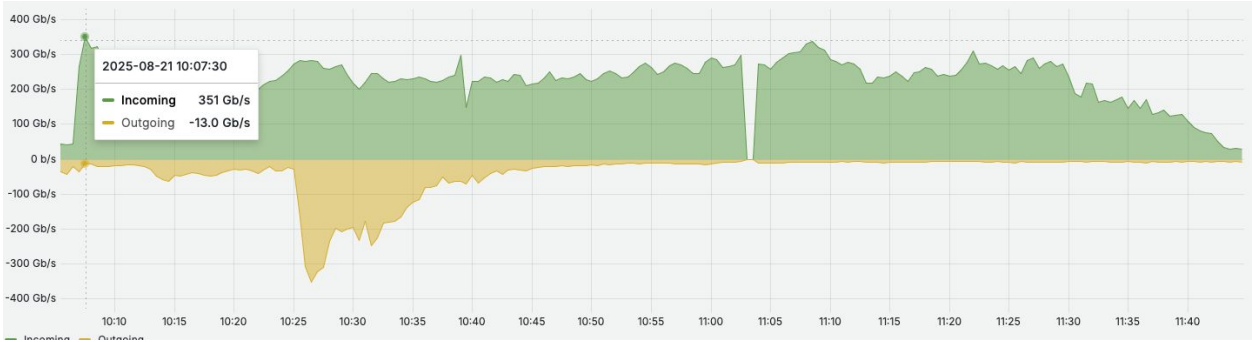
UCSD

- Recently upgraded their Network connectivity, all the way to ESnet, to 400 Gbps.
- Bottleneck is now at the DTN level. Didn't have enough time to deploy the new servers (after the lengthy upgrade above)
- 10 usable DTNs, each with 1x10 Gbps NIC
- When used in full-duplex (read and write at the same time) we only get about 8 Gbps (tested with iperf3)
- XRootD normally gets about 80% of what iperf (or similar tools) get
- That give us a bottleneck of 64 Gbps (we saw 40-50 during the test)

USCMS Week 1: Story Time (4/6)

FNAL (Reads)

- The target rate for FNAL during this mini-challenge was set at 800 Gbps
- For the Read test we pulled from all the T2s
- Adding up all the T2 Avg results from individual tests we only get: 520 Gbps
- The test peaked at 350 Gbps and then went down to 200-300 Gbps.



- We hit the FTS limit of 1200 Active transfers to FNAL (FNAL's FTS)

USCMS Week 1: Story Time (5/6)

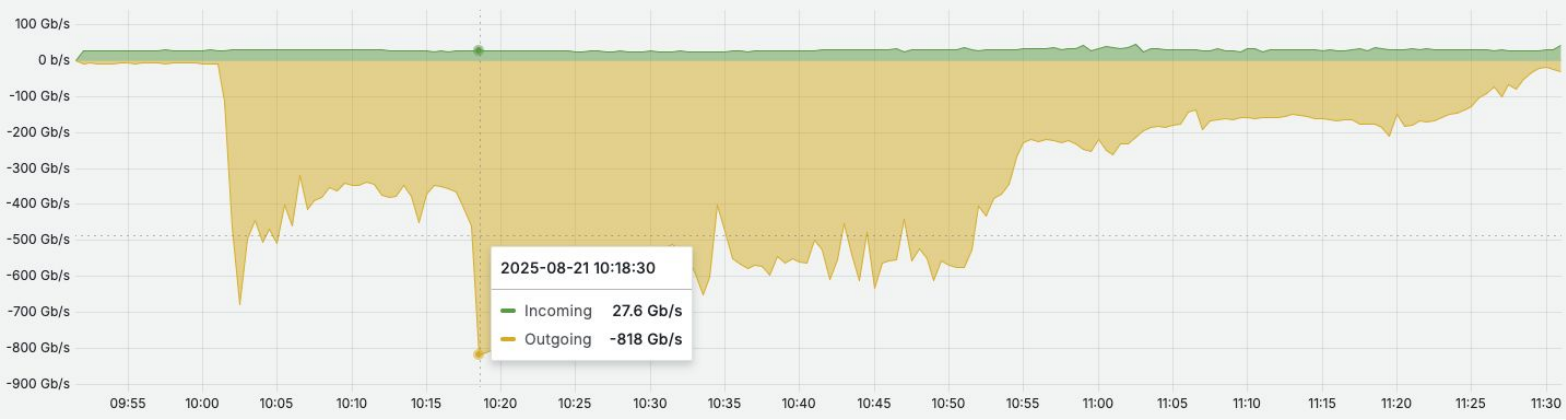
FNAL (Writes)

- During the write test we had to exclude Caltech due to issues with certs
- Adding up all the T2 Avg results from individual tests (minus Caltech) we only get: 605 Gbps
- By mistake we ended up using both fts3-cms and fts3-pilot servers at the same time, which significantly increase the overall number of active transfers to the point that FNAL refused to start new transfers

> *"INFO Thu, 21 Aug 2025 19:18:15 +0200; Davix: < HTTP/1.1 500 transfer not accepted: too many transfers!"*

USCMS Week 1: Story Time (6/6)

- On the other hand, we were able to hit our mark, even if just for a moment



- Sustained throughput was more like 600 Gbps
- The aggregated FTS limit of active transfers during this test was 3200

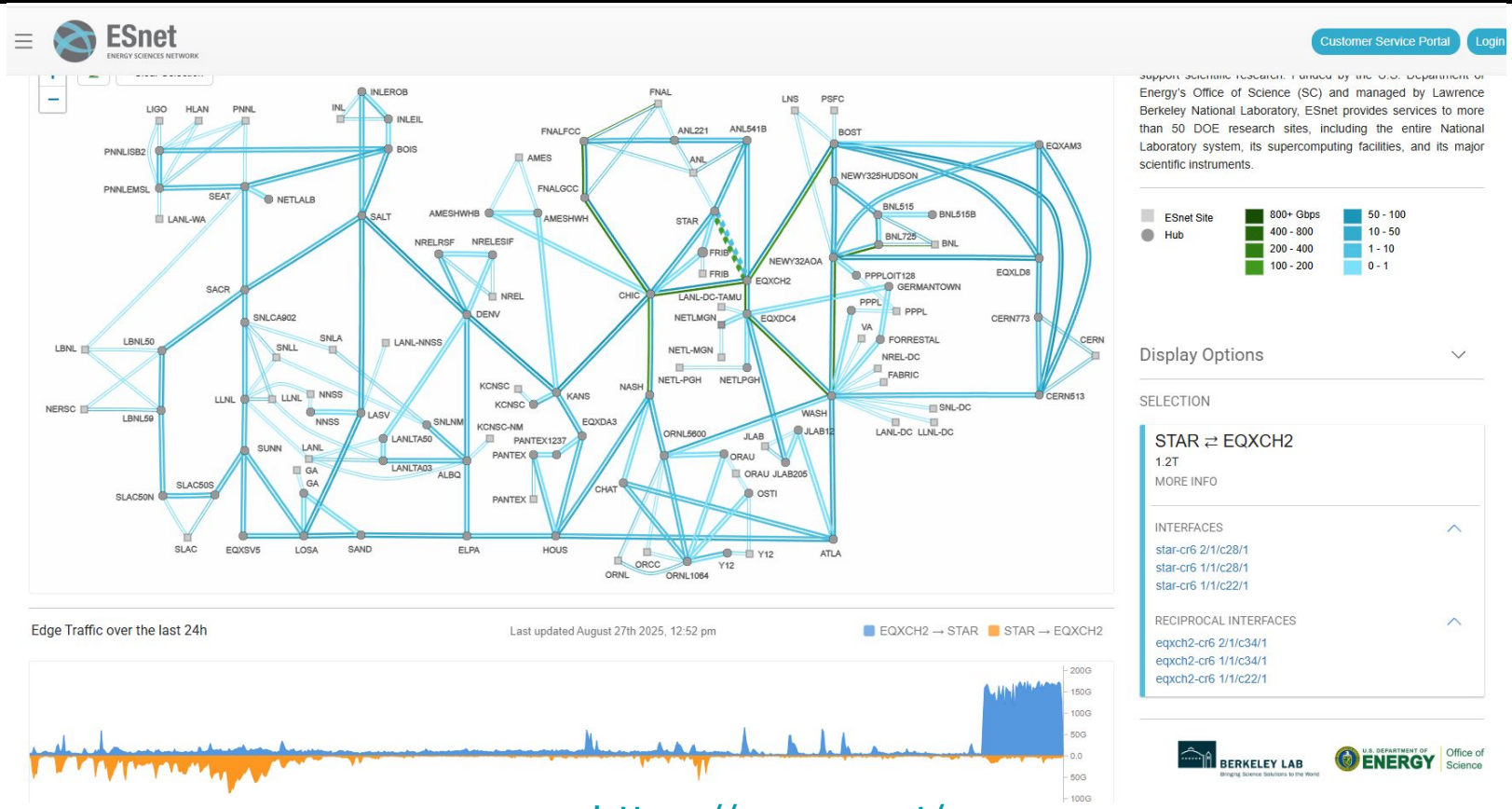
Schedule for Second Week: Joint Testing

The plan for the second week was to target jointly stressing USATLAS and USCMS sites.
We picked Wednesday August 27th from 10 AM to noon to run the joint test with August 28th as backup / re-do day
The plan:
USATLAS: BNL + NET2 => MWT2 + AGLT2 + SWT2
USCMS: FNAL => T2_US_*

We wanted to identify any hotspots in the shared network and tracert analysis gave us possible hotspots (the router location in ESnet in blue), sites in lists:
chic-cr6: ['caltech', 'nebraska', 'florida', 'ucsd', **'mwt2'**, 'caltech', 'nebraska', 'florida', 'ucsd', 'mit', 'vanderbilt', **'ou'**]
eqxch2-cr6: ['mwt2', **'mit'**, 'aglt2_msu', 'aglt2_um', 'mwt2_uc', 'ou']
fnalgcc-cr6: ['caltech', 'nebraska', 'florida', 'ucsd', 'caltech', 'nebraska', 'florida', 'ucsd', 'mit', 'vanderbilt', 'purdue']
wash-cr6: ['mwt2', 'swt2', 'aglt2_msu', 'aglt2_um', 'mwt2_uc', 'ou', 'swt2_cpb']
newy32aoa-cr6: ['mwt2', 'net2', 'swt2', 'aglt2_msu', 'aglt2_um', 'mwt2_uc', 'ou', 'swt2_cpb']

These locations are monitored in the my.es.net web page

ESnet Monitoring from August 27, 2025



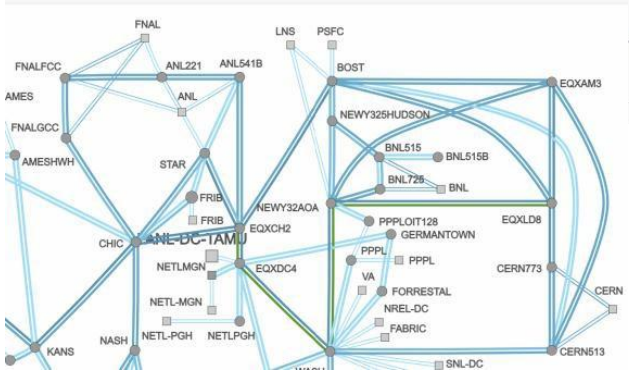
https://my.es.net/

Observation from Joint USATLAS/USCMS Test

Test: From BNL and NET2 to AGLT2, MWT2 and SWT2 (UTA) and FNAL to all USCMS Tier-2s

No obvious hotspots have been observed. The destination sites were obtaining data at their expected site rates.

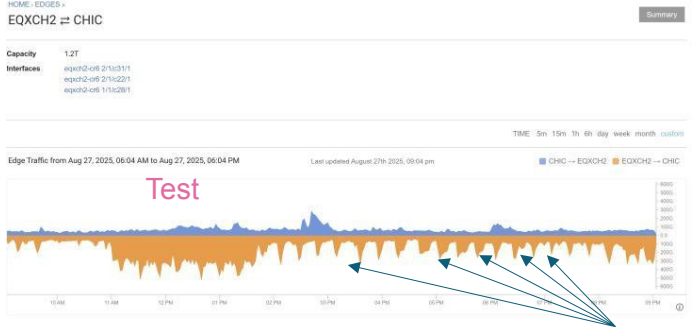
Periodic large data from CERN to FNAL is observed during test test although it didn't affect the data rate for USATLAS sites.



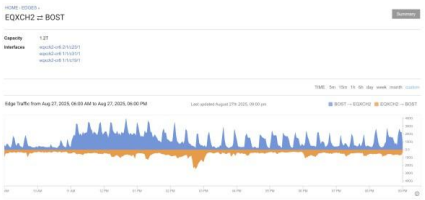
ESNET monitors

To AGLT2

To MWT2 and SWT2



Curious periodic traffic?



Joint Testing Results

From ATLAS: During the joint testing, we were able to reach similar rates per site compared to the prior week site testing results. There was no indication of network congestion or interference between USATLAS flows and other users of the network.

From CMS: The first day of the joint test we noticed a significantly lower rate than in the previous week. The following day the rate was basically the same as the previous week. On the first day we noticed an uneven distribution of active transfers across the T2s, this situation was improved the second day, hence we believe this was the reason for the discrepancy observed.

What Did We Learn?

USCMS

FNAL's FTS

- Found a misconfiguration on the overall limit of active transfers following an upgrade

Nebraska:

- Failure to include a /60 subnet for SENSE in Nebraska's CRIC NetworkRoute list resulted in ¾ traffic not traversing expected LHCONE path due to FNAL basing their routing policies on CRIC data. A misconfiguration caused this subnet to get used in the first place for more than just SENSE traffic.
- Significant rate of discarded packets between DTNs switch and ToR one. Fixed during the test by adding capacity

Purdue:

- It was discovered that Purdue still has a 2x100Gb limit between their CMS resources and the Purdue research network which had been assumed to be a 400Gb capable path.

Vanderbilt and Wisconsin

- They though they had 200 Gbps out of their T2, in practice they do not (needs further investigation)

What Did We Learn?

USATLAS (We would benefit from another test, grabbing more details)

- AGLT2
  - Identified broken WLCG site net monitoring at AGLT2_MSU (fixed)
  - Noted some dCache errors related to too large FTS settings during tests
- NET2
  - Was able to almost reach 400 Gbps (some tuning might max allow 400G)
  - Some evidence of CPU capping on dCache pools and headnodes
- MWT2
  - Reached 250 Gbps but was not fully able to exploit IU site, likely due to dCache pool selection. Tuning to be explored.
- SWT2
  - Both UTA and OU “DTN” bottlenecks (each has 3 10G servers for data)
  - Planning new hardware to exploit 100G network connectivity
- BNL still needs to upgrade to dual-homed pool nodes to fix transfer via doors

Summary & Plans

The US is continuing a series of mini-challenges for both capacity and capability.

This August 2025 capacity exercise has shown us a number of interesting results that help inform the experiments and the sites about issues to watch for and correct, as well as future directions to explore in terms of hardware, architecture and software…

We need to continue regularly running mini-challenges (both capacity and **capability**) in support of the WLCG Data Challenges.

- Is 3-4 times per year the right cadence?
- Capability effort? We need advocates/organizers!!

Questions or Discussion?

Acknowledgements

Thanks to the USATLAS and USCMS teams for carrying out the tests, especially Diego Davila and Hiro Ito from running the 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:

WLCG DOMA wiki page: https://twiki.cern.ch/twiki/bin/view/LCG/DomaMiniChallenges (includes link to Google Calendar to track activities)

Presentations

- 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)
- Upcoming Mini Challenges in the US (Nov 2024)
- Results form US Mini-Challenges (Jan 2025)
- US Mini Data Challenge Plans (Feb 2025)

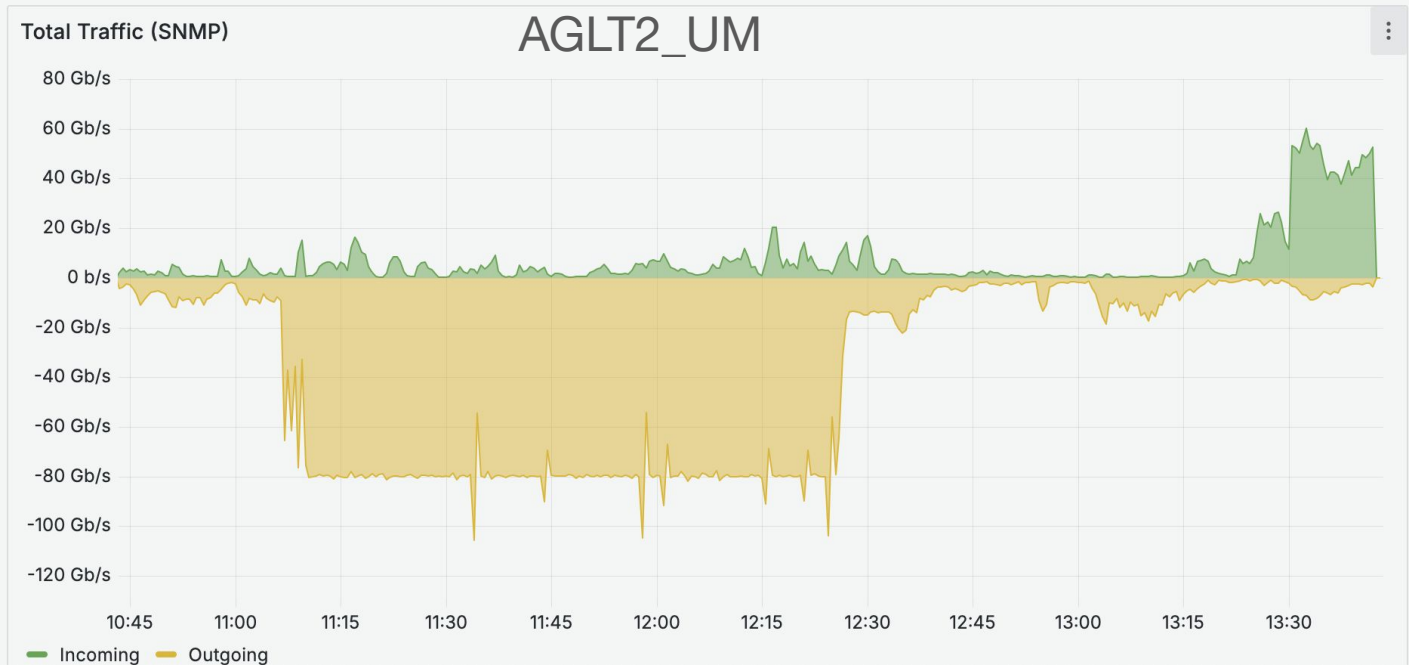
Some Google Docs

- WLCG/DOMA Data Challenge 2024: Final Report
- WLCG Mini-Capability Documents Folder
- 2025-08 US Capacity Mini-challenge

Backup/Detail Slides

USATLAS: AGLT2 Details

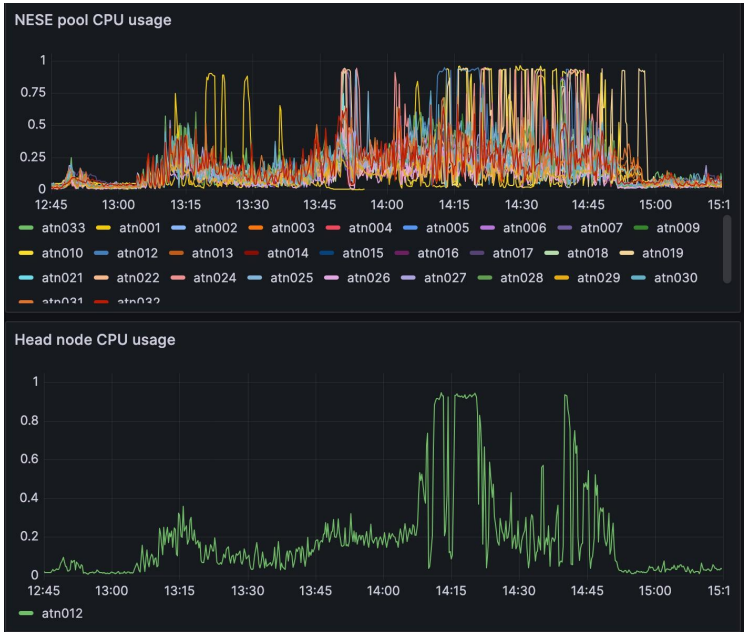
Site has a 180 Gbps limit (100 Gbps from MSU, 80 Gbps bottleneck at UM)
(Known issue. UM is waiting for new backbone switches proving 100 G ports)



USATLAS: NET2 Details

NET2 didn't have their 400G connectivity in fall 2024 and this capacity challenge was the first to use their 400G links (to ESnet in Boston and New York). Overall, no transfer errors/failures.

During this test, some CPU capping was Observed =>



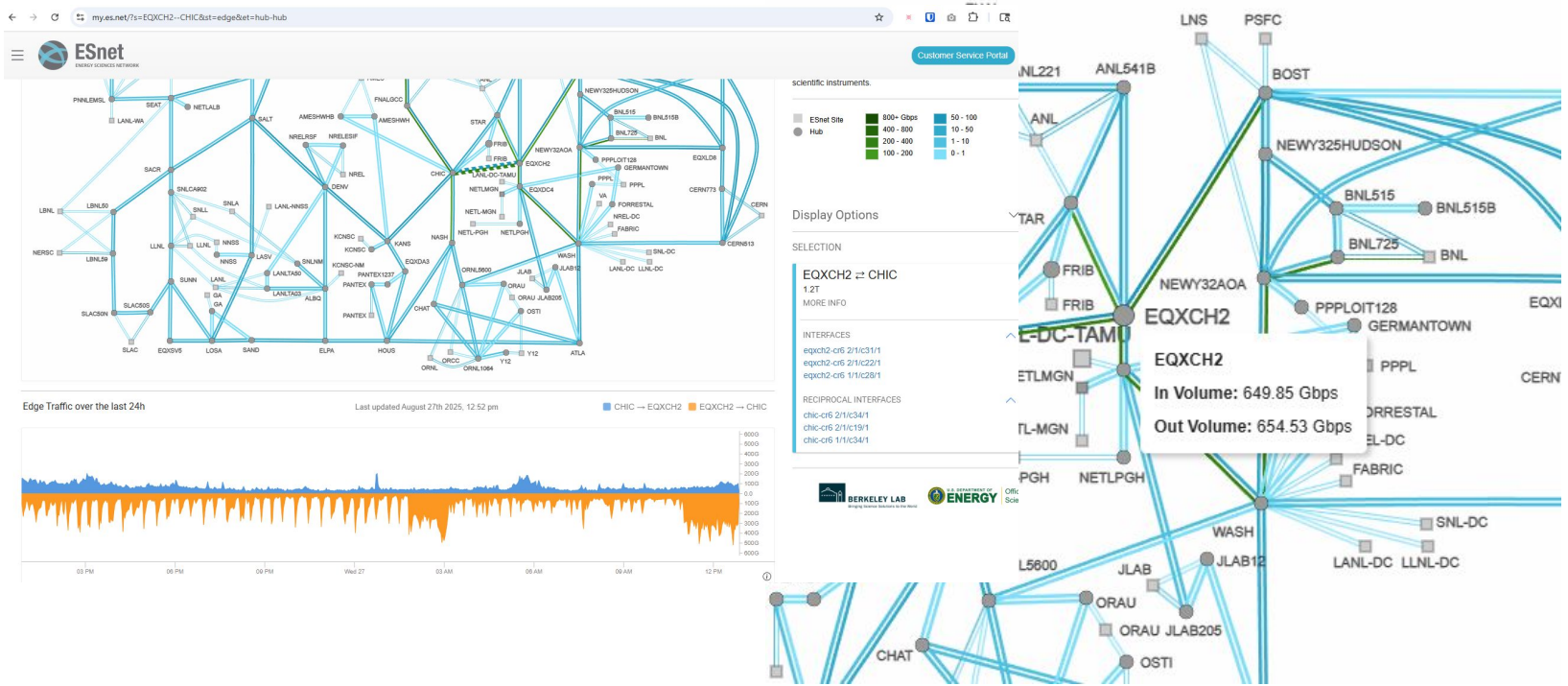
FTS efficiency (BNL=>T2 8/28) ~100%

Details for davs://dcgftp.usatlas.bnl.gov → davs://webdav.data.net2.mghpcc.org

Storage	Current active transfers	Max. active transfers
davs://dcgftp.usatlas.bnl.gov	105	9500
davs://webdav.data.net2.mghpcc.org	466	800

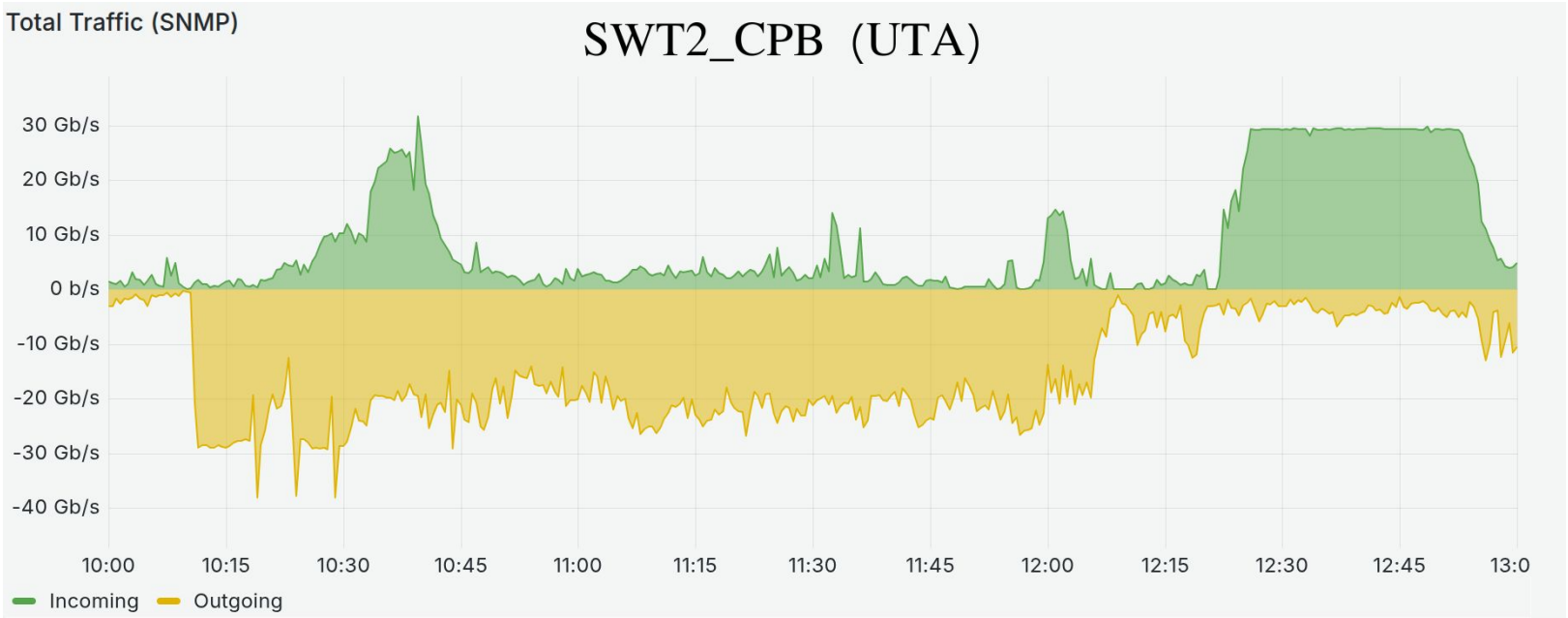
USATLAS: BNL Details

Need to get site report details

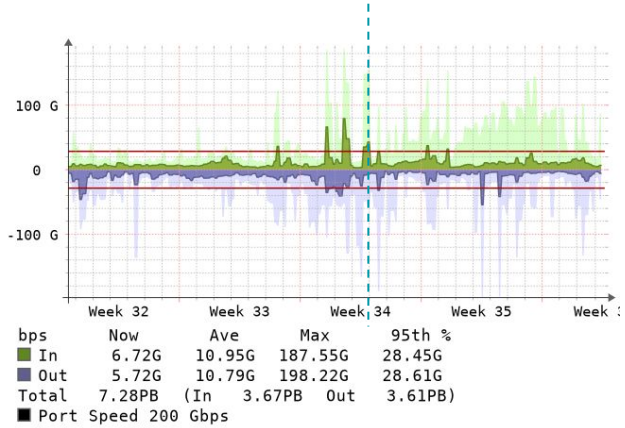


USATLAS: SWT2 Details

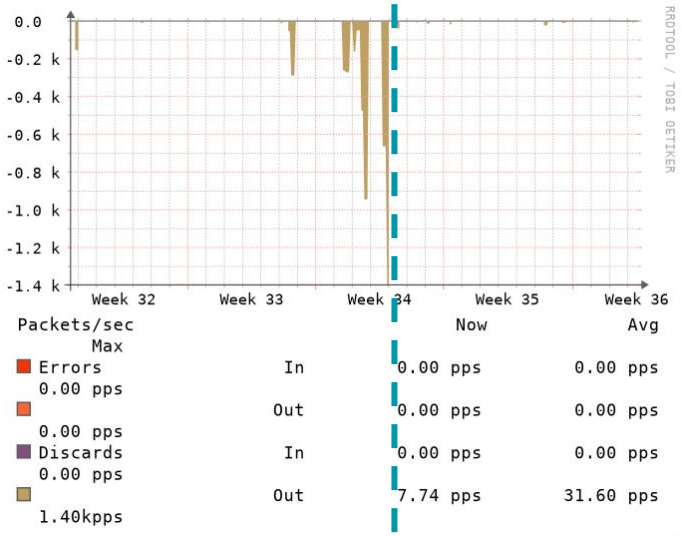
The current bandwidth is capped to 30 Gbps at both UTA and OU, limited by 3 10 G DTNs at each site. Replacement hardware being procured



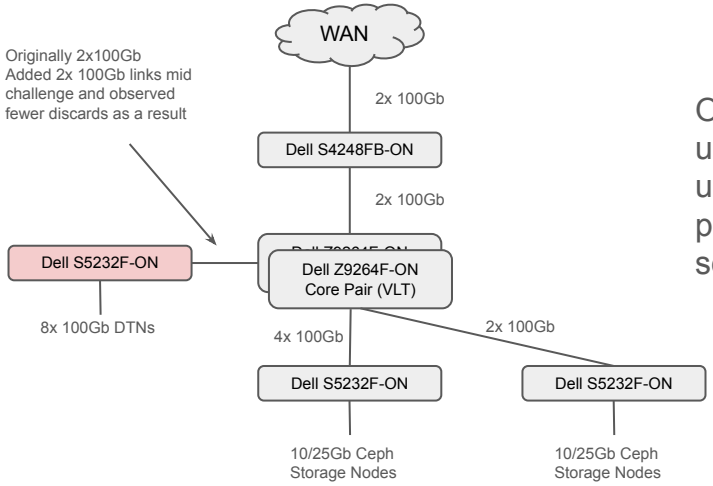
T2_UNL_Nebraska DTN ToR Discards



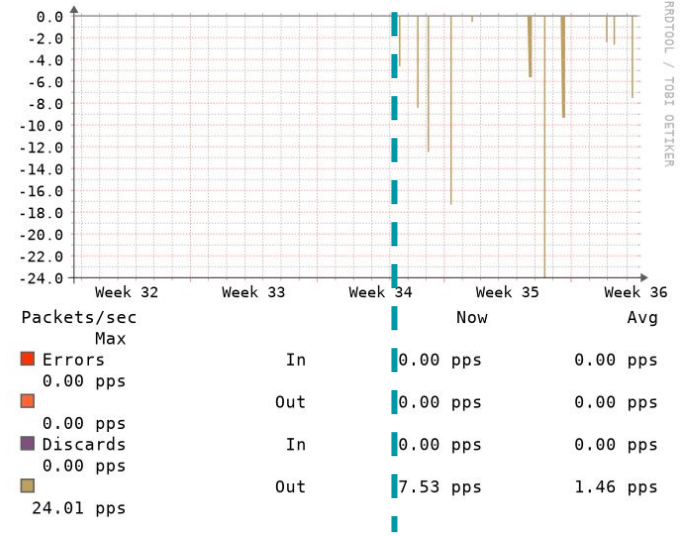
One of the 2x100Gb ToR uplinks on switch DTNs connected to during heavy traffic. Upwards of 1k discards per second observed.



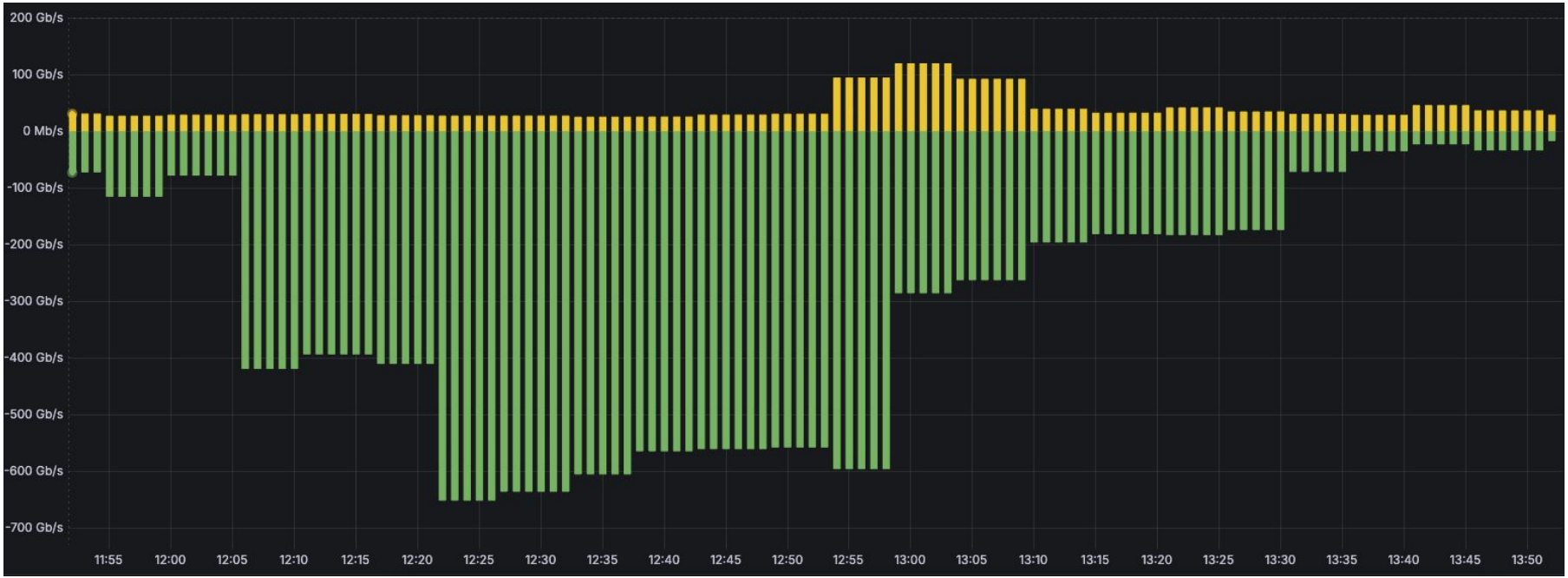
Dashed line is when additional 2x100Gb uplinks added to switch



One of the added 100Gb ToR uplinks. Discards on all 4x 100Gb uplinks with the same traffic pattern reduced to a dozen per second observed.

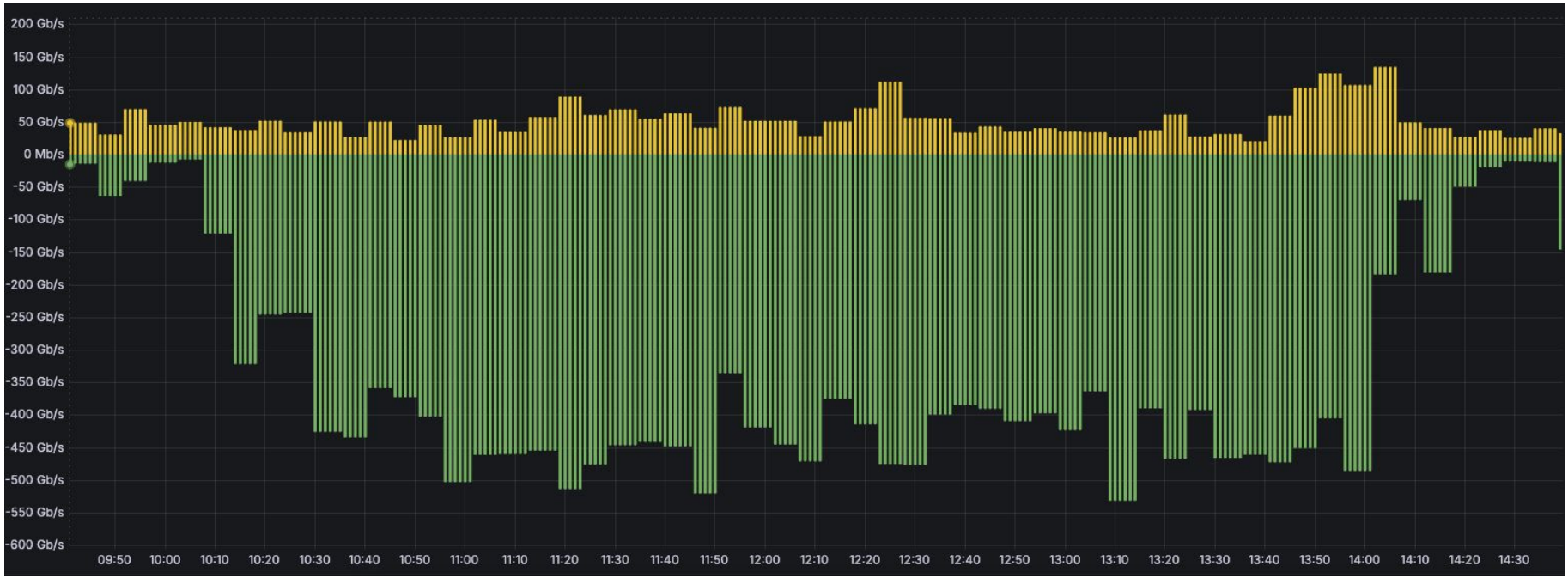


FNAL original test



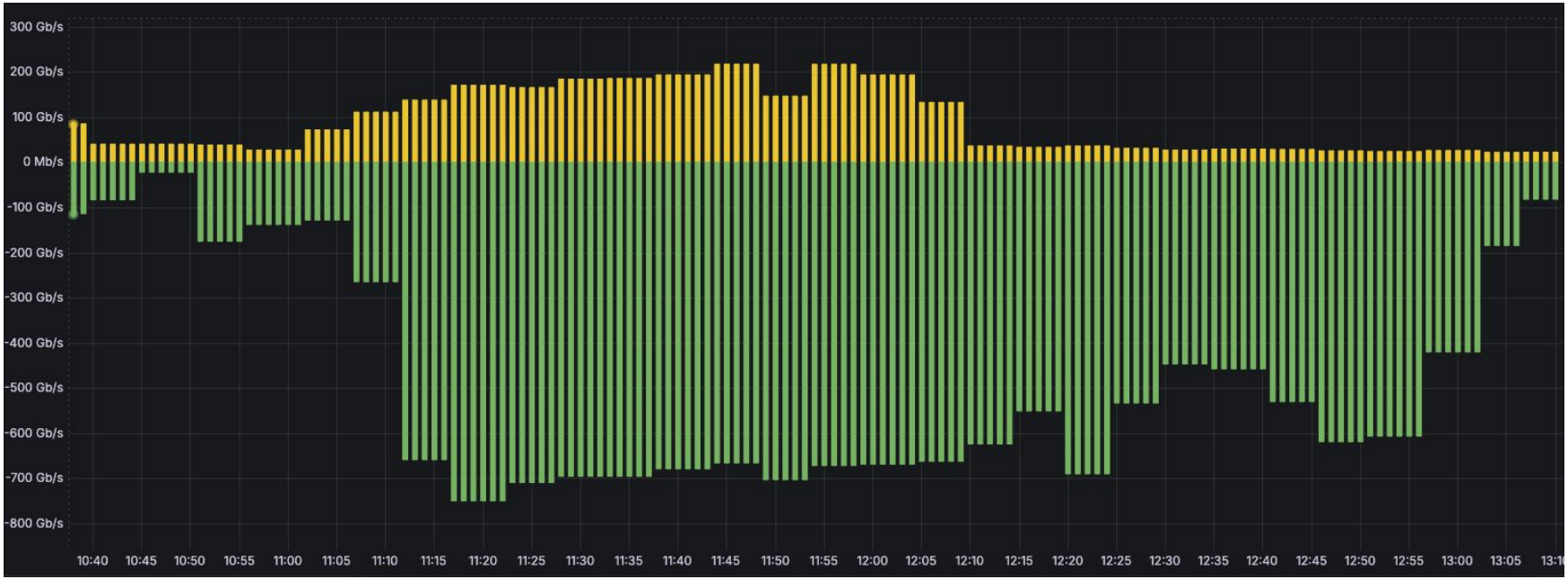
http://monit-grafana-open.cern.ch/d/MwuxgogIk/wlcg-site-network?orgId=16&from=2025-08-21T16:51:40.931Z&to=2025-08-21T18:52:54.167Z&var-bin=1m&timezone=America%2FChicago&var-site=USCMS-FNAL-WC1

Joint Test first day, FNAL's PoV



http://monit-grafana-open.cern.ch/d/MwuxgogIk/wlcg-site-network?orgId=16&from=2025-08-27T14:40:52.477Z&to=2025-08-27T19:39:36.609Z&var-bin=1m&timezone=America%2FChicago&var-site=USCMS-FNAL-WC1

Joint Test second day, FNAL's PoV



http://monit-grafana-open.cern.ch/d/MwuxgogIk/wlcg-site-network?orgId=16&from=2025-08-28T15:37:29.862Z&to=2025-08-28T18:10:46.230Z&var-bin=1m&timezone=America%2FChicago&var-site=USCMS-FNAL-WC1