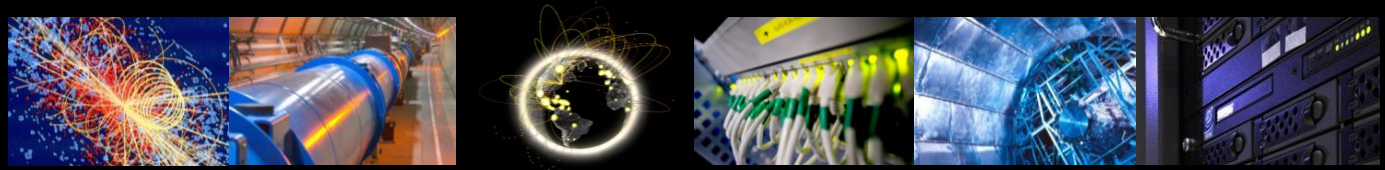
Upcoming Mini-Challenges in the US

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

WLCG DOMA General Meeting (https://indico.cern.ch/event/1470863/)

November 13, 2024



WLCG Worldwide LHC Computing Grid

Open Science Grid

WLCG Data Challenges

The **WLCG Data Challenges** are a ~biennial series of four increasingly-complex exercises which started in 2021 and are aimed at demonstrating readiness at the HL-LHC scale.

Next data challenge (DC26?) targets **50%** of HL-LHC scale and includes T1/T2 and any improvements we can integrate into our infrastructure.

These data challenges provide many benefits, allowing **sites**, **networks** and **experiments** to evaluate their progress, motivate and validate their developments in hardware and software and show readiness of technologies at suitable scale.

For **USLHC**, we believe it is critical to fully participate in future challenges, both by preparing and testing before each and analyzing the results after each.

What is a “mini-challenge”

Given the cadence and scope of the WLCG Network Data Challenges, it is important to understand what we mean by “mini-challenge”. This is our working definition:

- *A mini-challenge is a lightweight way to test capacity or capability of one or more sites*

The goal is to make it easy to test and track both our capabilities and capacities, finding and fixing bottlenecks, identifying bad architectures and hardware and improving our visibility into how our sites perform as part of a globally distributed infrastructure.

Transforming our Sites

The data challenges provide us with an opportunity to evaluate our existing hardware, software and architecture to identify bottlenecks, limitations and misconfigurations.

Given that HL-LHC is ~6 years away, now is the perfect time to re-evaluate our site's hardware configuration and architecture so that we can have a suitable baseline ready for HL-LHC requirements.

- Six years of hardware purchases can fully replace our current hardware
- Incrementally transforming sites should allow a smooth transition in capability

It is **critical** that sites understand how they fit into our globally distributed infrastructure so they can meet the HL-LHC requirements and use-cases.

- Mini-challenges are a great opportunity to understand our current capabilities, identify bottlenecks and prototype new technologies.

Goals for DC26/27

For DC26 (or DC27 if it moves later) we are targeting:

- All sites should be moving the majority of their data via **IPv6**
- We should have a few **IPv6-only** sites for each experiment
- At least 80% of the traffic should be **identified via SciTags**
- At least 50% of the traffic should be using **jumbo frames**
- **Rucio/SENSE** to be used by few Production sites
- Sites should be able to easily utilize **90% of their declared WAN bandwidth** for an extended period (many hours to days)
- **Network** traffic monitoring should be able to track throughput by network type (LHCOPN, LHCONE, Research & Education, Commercial/Commodity) organized by the WLCG Monitoring Task Force

Mini-challenges should help get us there…

USATLAS and USCMS Activities: What is in Common?

If we are going to undertake various mini-challenges that are experiment-based, there may be benefit in running them at the same time.

Certain resources or network transit locations may be shared and identifying any contention can be an important outcome from running tests at the same time so we can address the issues as they are observed.

On the other hand, we don't want to make mini-challenges heavy weight or over constrained as to when they be run. One of the advantages of mini-challenges are supposed to be the ability to quickly and easily test capability or capacity for a site or set of sites.

Management and Logistics for Mini-Challenges

Commonly mini-challenges that test **new technologies** are run independently, (specially if they do not interfere with day-by-day Ops) by the people around the related project

On the other hand, those that run throughput tests are run in coordination with the site admins of the sites involved and the relevant R&E network responsibles.

In any case every mini-challenge should be reported in advance to the WLCG DOMA conveners. This presentation is serving as notice of mini-challenges planned for the US but we should have **guidance** about the best practice here.

Running periodic 1-to-1 load tests (NANO-challenge?)

We want to regularly “benchmark” our facilities and networks by:

- Running periodic throughput tests between the T1 and each T2:
  - Start by **clearly defining the target rates** for each site for DC26
  - Pick a week in the calendar that is suitable for both sites
  - Define a target throughput for the test
  - Use the *dc_inject* tool to execute the test

- Running periodic throughput tests for the T1
  - Run a similar test (as above) using a group of T2s to inject data to/from the T1

This requires each participant site to have a **validated** site network monitoring

Q. Who will do this?

A. for USCMS: Diego Davila (UCSD), for USATLAS: Hiro Ito (BNL)

Load Test Tool Overview

- Simple program to generate the network load was created prior to DC24
- The detail of the program was presented at various meeting. It can be found at Easy-to-use Network load generator and test results at USATLAS
- The program will transfer disk resident data from the desired source to target destination using preferred FTS service.
  - The throughput rate can be specified
  - The list of source files can be given as a list in a file or let the program to pick files from your preferred storage directory.
  - The program is found at the following BNLBox folder https://bnlbox.sdcc.bnl.gov/index.php/s/XGs6LJEGNzf69zK

USATLAS Mini Data Challenge Fall 2024 (1 of 2)

- 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,...

USATLAS Schedule of Load Tests (2 of 2)

USATLAS Schedule for load testing

Site	Date	Monitoring
AGLT2 (MSU/UM)	Nov 13	Yes
MWT2(IU/UC/UIUC)	Nov 19	Yes
NET2 (Amherst)	Dec 3	Yes
SWT2 (UTA/OU)	Nov 21	Not Ready
BNL	Dec 5	Yes
Multisite	Week of Dec 9	?
Tape (BNL)	Week of Dec 16	Yes

We want to get a new baseline of where our sites are at with a series of load tests.

Each **Tier-2** will have a read and write test from the **Tier-1** to identify its “maximum” with the current hardware/software.

The previous slide described the various types of tests planned.

USCMS DC26 mini-challenges Plans for Fall (1 of 4)

1. Validate all sites reporting to the WLCG monitoring dashboard:
https://monit-grafana-open.cern.ch/d/Mwuxgoglk/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
- FNAL: ~400 Gbps
- We can increase if Sites are ready to push harder

USCMS DC26 mini-challenges Plans for Fall (2 of 4)

Status: WLCG Monitoring Validation

SIte	Monitoring
Caltech	Yes
Florida	Yes
MIT	No
Nebraska	Yes
Purdue	yes
UCSD	broken
Vanderbilt	No
Wisconsin	Yes
FNAL	Yes

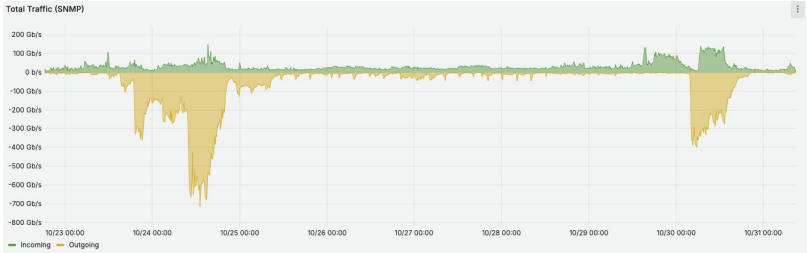
Information on how to set this up:
https://gitlab.cern.ch/wlcg-doma/site-network-information/

For container-based setup:
https://github.com/syedasifraza/site-net-monitor

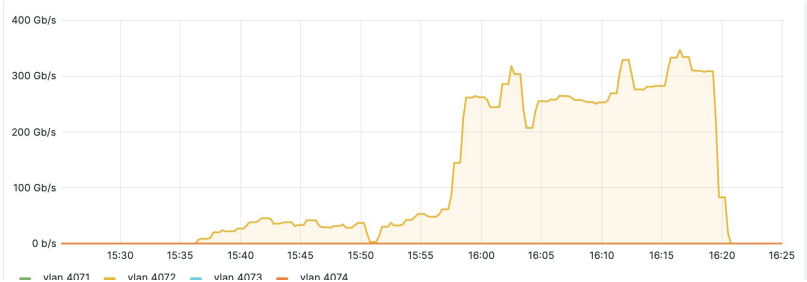
USCMS DC26 mini-challenges Plans for Fall (3 of 4)

Status: Load Tests

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 is ongoing a network upgrade :(

USCMS DC26 mini-challenges Plans for Fall (4 of 4)

Schedule: who wants to go first?

SIte	half-week
Florida	
MIT	
Nebraska	Nov 25-29 (1st)
Purdue	Dec 16-20 (1st)
Vanderbilt	
Wisconsin	

Available half-weeks	Free	Conflict
Nov 11-15 (2nd)	yes	Heavy ion(HI)
Nov 17-22 (1st)	yes	HI/SC24
Nov 17-22 (2nd)	yes	HI/SC24
Dec 2-6 (1st)	yes	
Dec 2-6 (2nd)	yes	
Dec 9-13 (1st)	yes	UK mini-challenge
Dec 9-13 (2nd)	yes	UK - FNAL mini-challenge
Dec 16-20 (2nd)	yes	Diego on vacation

Site admins: please contact me with your desired half-week
didavila@ucsd.edu

Site Testing, Motivation and Benefits

We hope to have sites as **enthusiastic participants** in planning and executing various mini-challenges.

This is a great opportunity to identify how each site performs and identify where there are issues.

It should be a significant help in defining how the site should evolve their hardware, software and architecture.

Preparing Technologies and Capabilities

HL-LHC will require more resources than we can currently afford.

- To address this, the experiments are working hard to optimize workflows
- New technologies and capabilities will play a critical role in bridging the gap

The WLCG data challenges are designed to regularly test where we are relative to where we need to be for HL-LHC.

Possible technologies to test and, if beneficial, integrate

- **New / improved storage servers** (Gen5 PCIe, NVMe, new NICs, etc)
  - Define/document LHC server best practice for hardware and configuration (like Fasterdata)
- **SciTags** (traffic identification anywhere in the network)
- **Traffic optimization** (via Jumbo Frames, pacing, new protocols)
- **Network Orchestration** (SENSE/Rucio, NOTED, GNA-g efforts, etc)
- Improvements (alternatives) to **WebDAV and Xrootd protocols (**RNTWG**)**
- Improvements to **storage elements** (dCache, Xrootd, STORM, EOS, etc)
- Evolution of **Distributed Data Management** (Rucio, FTS, etc)

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
- Improved site network monitoring to traffic traffic by LHCONE, LHCOPN, R&E and commodity

The plan:

- Before the end of 2024 rerun capacity tests for US sites to determine current values
- February 2025, execute a joint USATLAS-USCMS **capacity** mini-challenge for North America (identify current throughput limits by site and in aggregate)
- Early-to-mid Summer 2025, execute a joint USATLAS-USCMS **capability** mini-challenge for North America (storage tokens, traffic shaping, SciTags, SDN, etc.)

Summary

We have some concrete plans for USLHC testing for DC26/27

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

We have an **opportunity** to leverage DC24 and newer 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**, **Justas Balcas** and **Asif Shaw** for their contributions to the slides

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

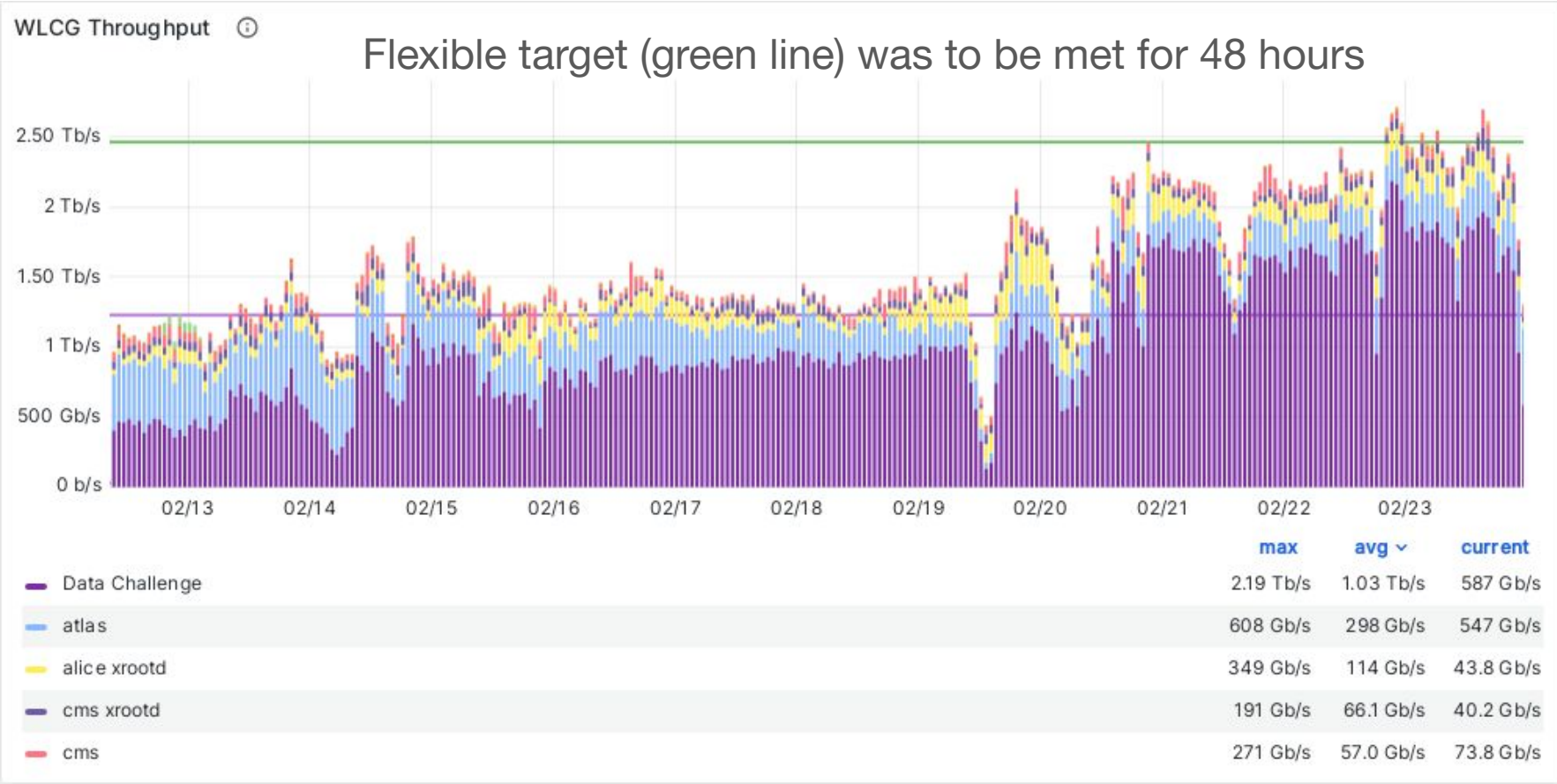
Katy Ellis LHCONE/LHCOPN DC24 presentation:
https://docs.google.com/presentation/d/1Tm3pCMkfHj5KHTW3PXbgS7mdHf72Ir27qr1JgMbrnRg/edit#slide=id.g1ea89411ecb_0_4

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

DC24 Throughput

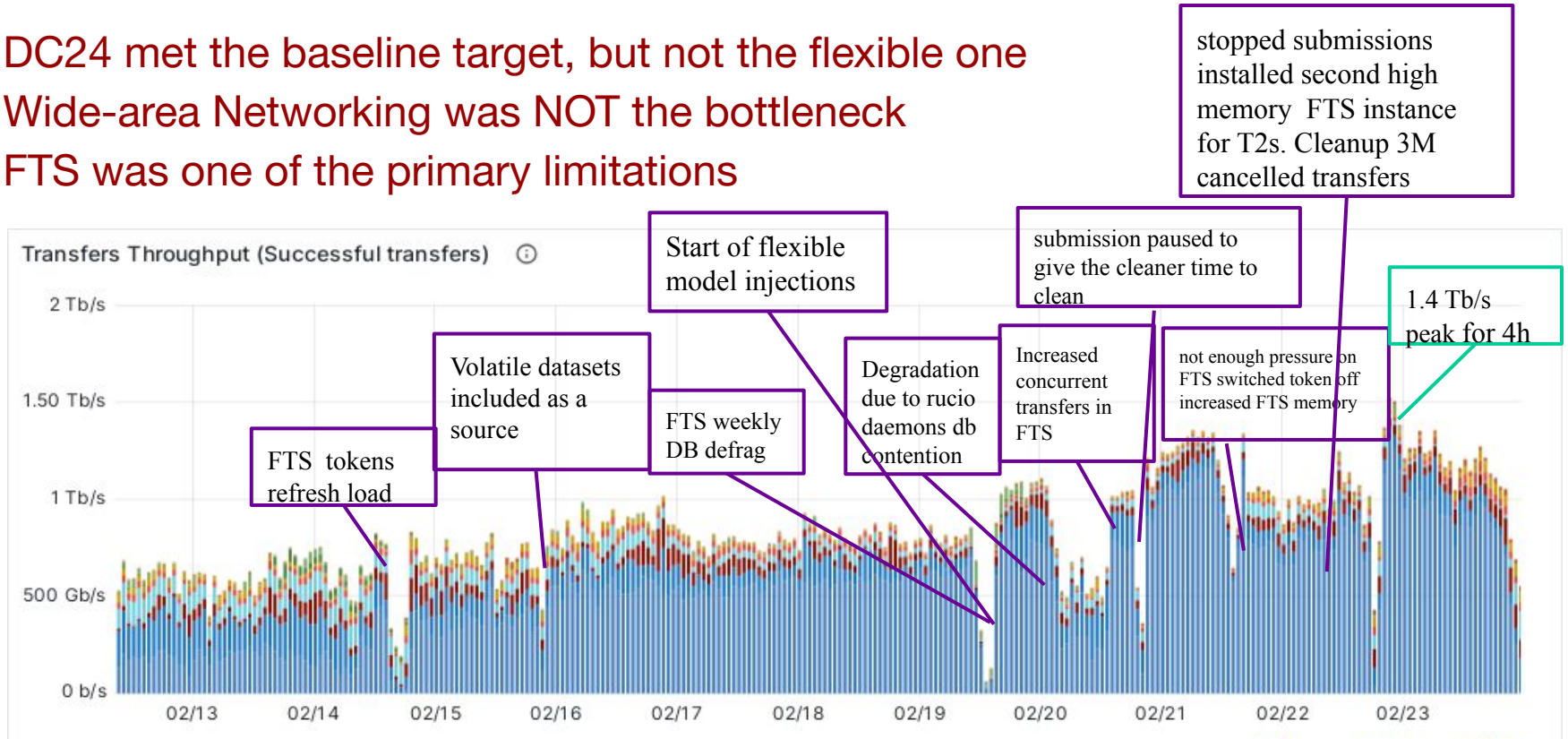


DC24 Throughput Annotations

DC24 met the baseline target, but not the flexible one

Wide-area Networking was NOT the bottleneck

FTS was one of the primary limitations



USCMS planned upgrades

- Upgrade FNAL to 1.6 Tbps total link capacity
- Jumbo Frames performance evaluation
- FNAL working on distributed load generation technique to test Terabit science networks (as opposed to use the injection link)
- Develop a Perfsonar's Alert & Alarm (Grafana Dashboard)
- Deploying Flow label & Packet marking techniques
- Working with ESnet on AI/ML based traffic classification
- Move SENSE/Rucio into pre-production: add more sites, include ATLAS

Needed Visibility for Data Challenges

For DC24, a site network monitoring campaign was undertaken to provide better visibility into each site's capabilities (see CERN Gitlab)

- This was a result of DC21 noting a deficiency in our monitoring
- For DC24 we just wanted total IN/OUT for each site
- We still have USLHC sites **missing** (see plot)
- For DC26, we may want to improve the level of detail (traffic by experiment)

We need to continue to improve (and verify) the monitoring we have, since this underlies all our attempts to identify friction points in our infrastructure.

For DC26, we would like to have at least 80% of our WAN traffic identified via the SciTags Initiative (there is also an IRIS-HEP metric for number of USLHC sites marking traffic; currently the number is '1' [Nebraska])

We need to identify what monitoring is missing and fix any incorrect monitoring and clarify any misleading monitoring.