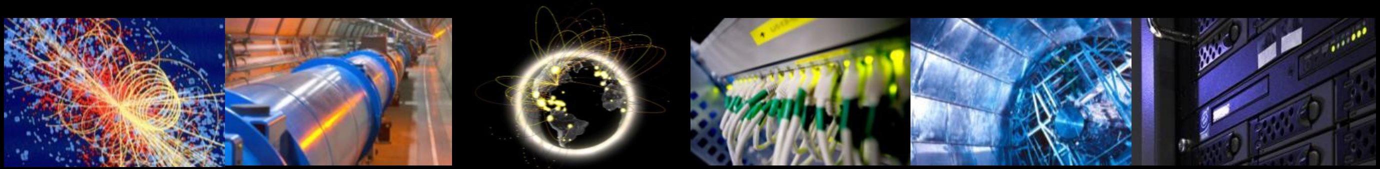
Post-DC24 Activities and Plans for DC26

Shawn McKee / University of Michigan & Farrukh Khan / FNAL

HTC 24

(https://agenda.hep.wisc.edu/event/2175/sessions/3176/#20240710)

July 10, 2024



WLCG Worldwide LHC Computing Grid

Open Science Grid

WLCG Data Challenges

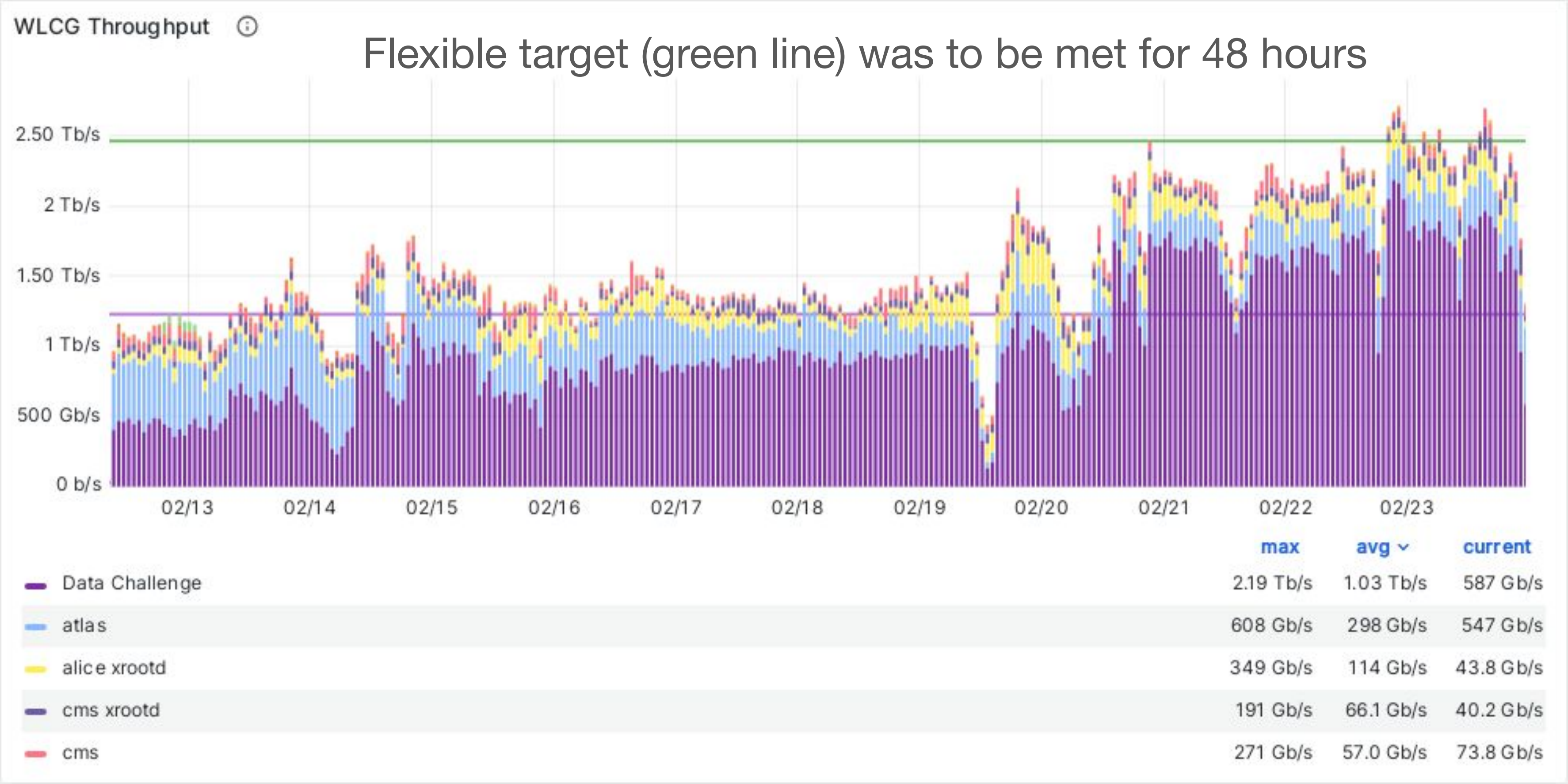
The **WLCG Data Challenges** are a ~biennial series of four increasingly-complex exercises which started in 2021 and are aimed at demonstrating readiness at the HL-LHC scale.

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.

DC24 Throughput



Flexible target (green line) was to be met for 48 hours

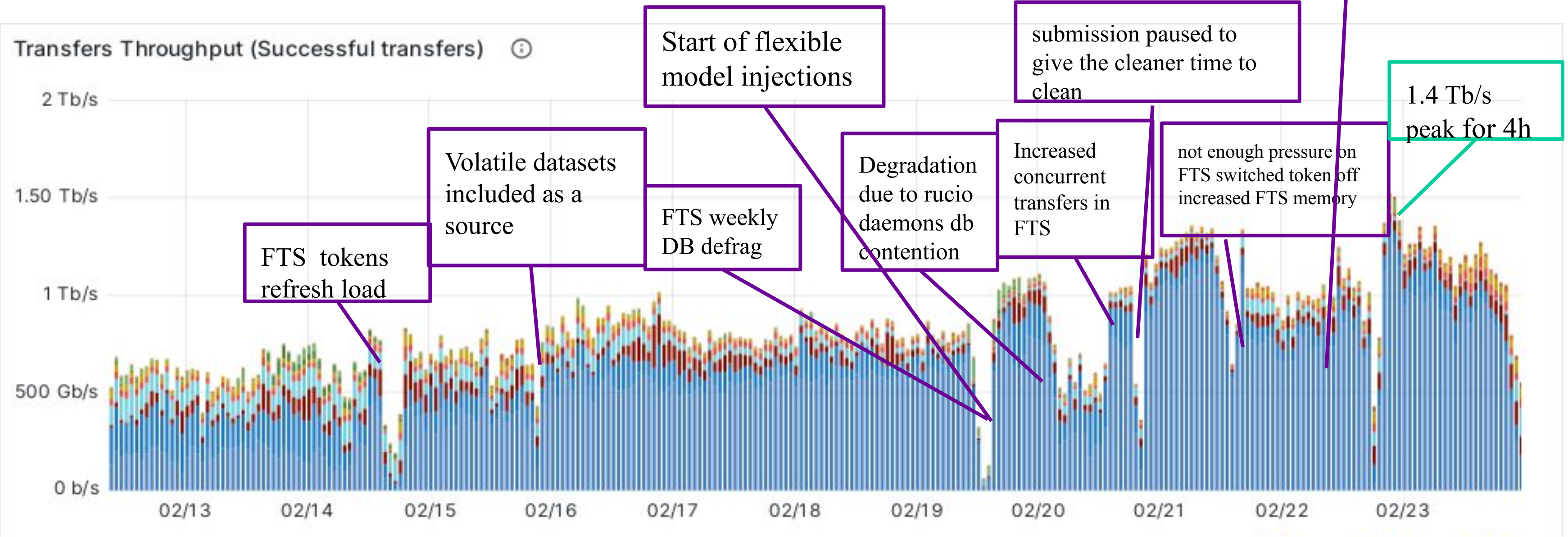
	max	avg ⌄	current
Data Challenge	2.19 Tb/s	1.03 Tb/s	587 Gb/s
atlas	608 Gb/s	298 Gb/s	547 Gb/s
alice xrootd	349 Gb/s	114 Gb/s	43.8 Gb/s
cms xrootd	191 Gb/s	66.1 Gb/s	40.2 Gb/s
cms	271 Gb/s	57.0 Gb/s	73.8 Gb/s

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



stopped submissions installed second high memory FTS instance for T2s. Cleanup 3M cancelled transfers

Start of flexible model injections

submission paused to give the cleaner time to clean

1.4 Tb/s peak for 4h

Volatile datasets included as a source

FTS weekly DB defrag

Degradation due to rucio daemons db contention

Increased concurrent transfers in FTS

not enough pressure on FTS switched token off increased FTS memory

FTS tokens refresh load

Goals for DC26

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 available 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)

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.

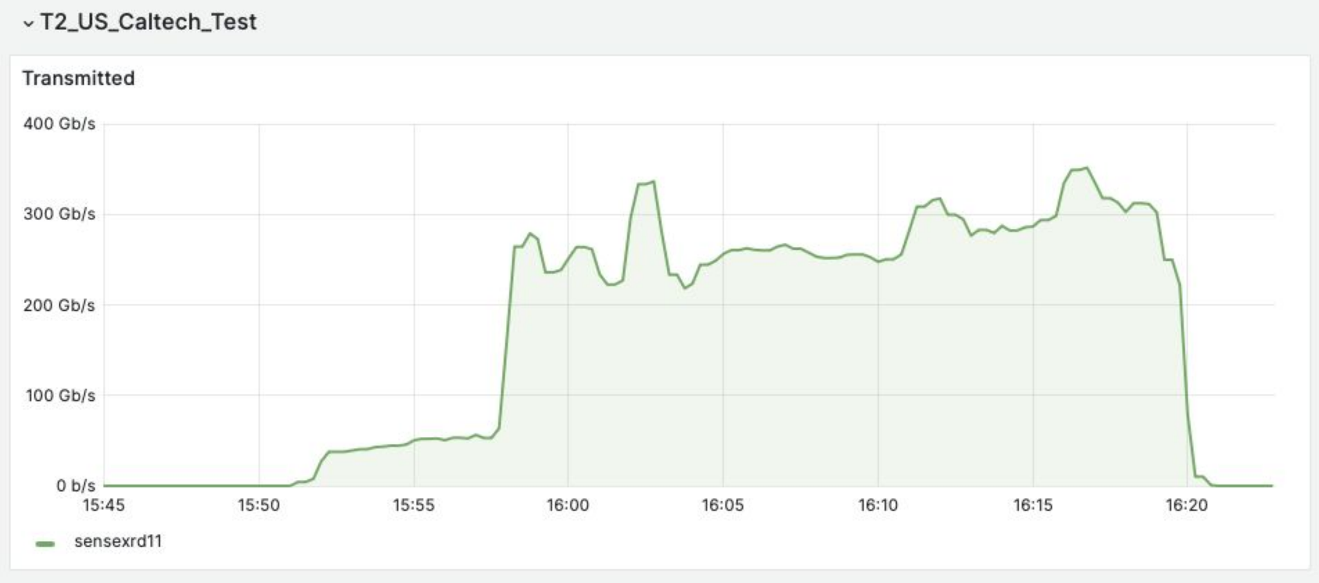
- Six years of hardware purchases can fully replace our current hardware
- Incrementally transforming sites should allow a smooth transition in capability

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.

Transforming our Sites (cont'd)

An example of the above is the case of Caltech that went from performing at 80Gbps during DC24 to ~250Gbps this week by analyzing the results from DC24 and upgrading and tuning their system



Caltech (Prod) => UCSD (test) HTTP-TPC transfer throughput

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.

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
- **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**
- Improvements to **storage elements** (dCache, Xrootd, STORM, EOS, etc)
- Evolution of **Distributed Data Management** (Rucio, FTS, etc)

Mini-Challenges and Ongoing Testing

The mini-challenges prior to DC24 turned out to be very beneficial for finding problems in our infrastructure and we should plan to have regular mini-challenges going forward

- The injection tool, used for DC24, is relatively easy to use. We could/should use it to run periodic tests with individual sites e.g. FNAL => MIT to help them understand possible limitations
- Hiro Ito / BNL has developed a load tester that can also be used
- Regional sets of sites (up to "North America") should be tested simultaneously by USATLAS/USCMS to verify we don't conflict at PoPs
- Mini-challenges should also include tests of technologies and new capabilities

Ongoing mini-challenges a few times a year provide important guidance and validation for site changes in hardware, software and tunings.

Questions To Discuss (and Try to Answer)

How do we best organize mini-challenges in common between USATLAS/USCMS?

- Use the HSF calendar to identify upcoming tests?
- Use Google docs to develop and define tests?
- How best to coordinate with ATLAS/CMS?
- What cadence is best: 1, 2 or 3 times a year?

Who will organize and operate the tests?

Who will analyse results and identify bottlenecks?

What technology-focus mini-challenges are needed?

- Driven by advocates or site interest?
- Co-scheduled with capacity mini-challenges?

How do sites share and implement beneficial changes identified by mini-challenges?

Summary

We (USLHC) need to **clarify** and **document** existing plans, mini-challenges and goals for the next year and for DC26/DC27

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

Question, Comments, Discussion?

Acknowledgements

Thanks to Justas Balcas, Diego Davila 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

Backup Slides

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

UCSMS DC24 summary

The planned target for FNAL was 32GB/s
The maximum injected/achieved rates for FNAL were 26/31GB/s for reads and 37/39 GB/s for writes making this a success for the T1.
Moreover, sustained rates of 24 GB/s for writes and 36 GB/s for reads were observed for 21 and 12 hrs. respectively

From the T2s perspective the initial targets set by CMS centrally, ranging from 1.2 to 4.8 GB/s, were dimmed lower than expected by USCMS and all sites performed above them.

On the last day, these rates were significantly increased as per USCMS request. The new rates ranged from 4 to 14 GB/s. This time only Nebraska performed above the new rates with other sites hitting marks around 50% of the updated targets.

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