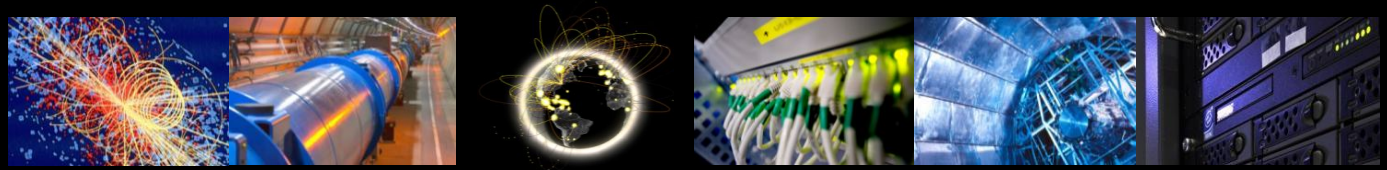
ATLAS Network Data Challenge: Plans for DC27

Shawn McKee / University of Michigan

6th Global Research Platform Meeting

(https://grpworkshop2025.theglobalresearchplatform.net/)

September 15, 2025



WLCG Worldwide LHC Computing Grid

Open Science Grid

Introduction

I am a Research Scientist at the University of Michigan Physics Department working on the ATLAS project.

My roles within ATLAS include being the Distributed Data Management co-coordinator as well as the USATLAS Facilities and Distributed Computing co-manager.

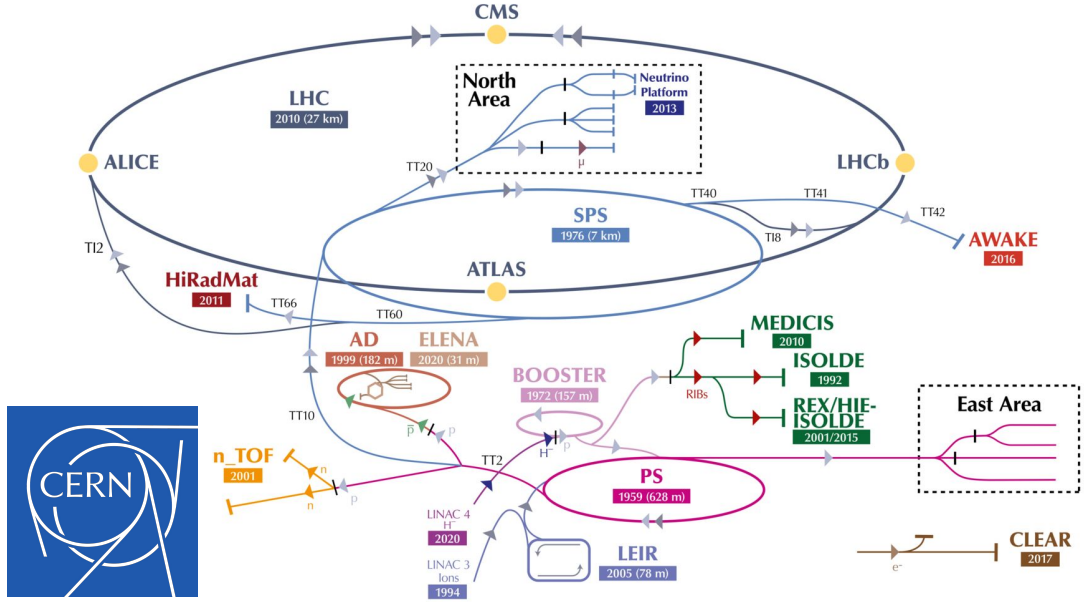
This presentation represents **my** ATLAS perspective on the WLCG Data Challenges but may not represent official ATLAS perspective (I am speaking from my own experience working on and preparing for data challenges for ATLAS).

The content reviews the basics and provides an update of what was presented in the 5th GRP.

Please feel free to ask questions as I go or save them up for the end of the talk.

The LHC, the WLCG and ATLAS

For those not familiar with High-Energy Physics, a quick introduction:



Shown in the figure are some detailed components of the LHC as well as the locations of the 4 main LHC experiments

The **L**arge **H**adron **C**ollider (**LHC**) at the European Laboratory for Particle Physics (CERN https://home.cern/) is the most powerful particle accelerator in the world. Highly energetic protons, traveling almost at the speed of light around a 27 kilometer long ring in both directions, are steered to collide head-on, creating new particles and new interactions to probe fundamental natural laws.

The WLCG and ATLAS Experiment

The **W**orldwide **L**HC **C**omputing **G**rid (**WLCG**) (https://wlcg.web.cern.ch/) collects resources worldwide and enables their usage by the LHC experiments as a distributed computing facility. WLCG is co-ordinated by CERN. WLCG is managed and operated by a worldwide collaboration between the experiments (ALICE, ATLAS, CMS and LHCb) and partners with EGI (European Grid Infrastructure), OSG (Open Science Grid), and NeIC (Nordic e-Infrastructure Collaboration).

ATLAS (http://atlas.cern/) is the largest of four particle detectors that measure and record the particle collisions at the LHC. The primary scientific goal is to quantitatively measure and discover properties of the Standard Model (SM) of particle physics.

ATLAS and LHC Running Organization

The LHC schedule since its startup is organized into long periods of particle collisions (called **Runs**) and periods of repair/maintenance (called Shutdowns). Between Runs are **Long Shutdowns** for upgrades and replacements

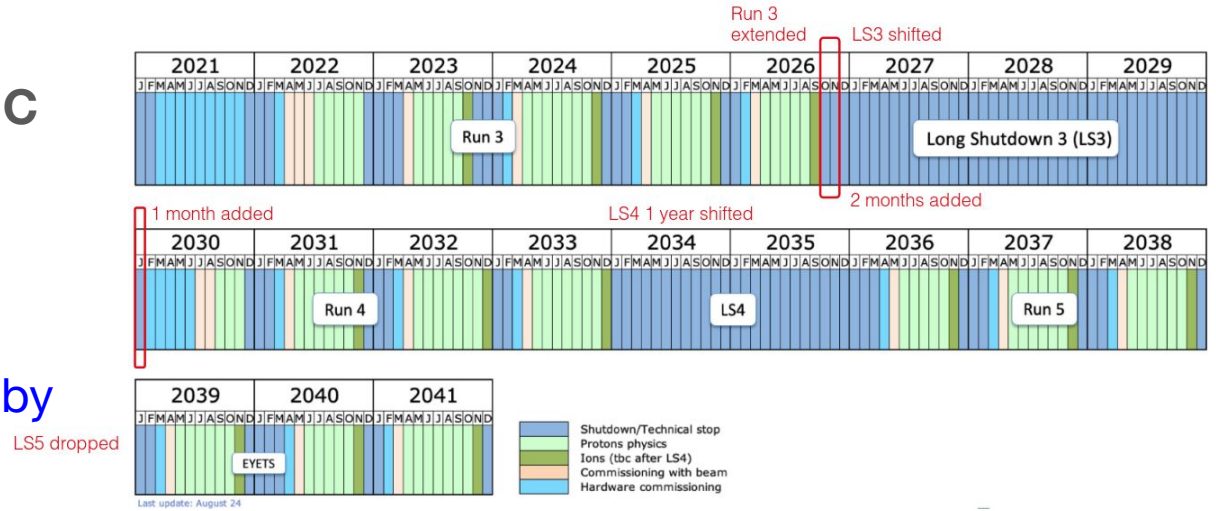
A revised schedule is under discussion which may extend the current **Run-3** and modify **HL-LHC** running in 2030

Waiting to see if official schedule changes (perhaps by Dec 2025?)

Update on LS3 schedule discussions

ATLAS EXPERIMENT

Mike Lamont's **preliminary** "Variant 3"



High-Energy Physics Resource Challenges

The WLCG experiments generate **~200 Petabytes** of raw data yearly from their detectors and additionally generate a similar volume of simulated and transformed data, which must be tracked, managed and made globally available.

- Since WLCG resources (computing, storage, services, etc) are also globally distributed, **networking becomes a critical component** for being able to pursue the scientific goals of the experiments.
- Additionally the resource requirements, in terms of the amount and type of computing accessible and the volume of storage, continue to **increase** while budgets remain (so far) **flat**.

Snapshot of CERN WLCG data animation

This means WLCG experiments, like ATLAS, need to be **innovative** to do more with the resources we can afford and need continual development and improvements to meet our future needs.

WLCG has planned a series of **data challenges** to evaluate how our infrastructure is developing to meet future needs and focus effort on identified bottlenecks.

WLCG Data Challenges

The **WLCG Data Challenges** are a ~biennial series of four increasingly-complex exercises which started in 2021(**DC21**) and are aimed at demonstrating readiness for the High Luminosity (HL) LHC scale in 2030.

Next data challenge (**DC27**) targets **50%** of HL-LHC scale and includes T1/T2, tape systems and any improvements we can integrate into our infrastructure. (**DC24** from February 2024 was **25%** of HL-LHC)

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.

I believe it is critical for our sites and the WLCG experiments to fully participate in future challenges, both by preparing and testing before each and analyzing the results after each.

What Have We Learned from Data Challenges?

The two WLCG Network Data Challenges we have run so far have been very helpful for the experiments, sites, R&E networks and technology innovators.

- **Sites** were able to identify bottlenecks that were not obvious before
- **R&E networks** were able to gain understanding about how the various participants data flows might interact with each other across their topology.
- The **WLCG experiments** and partners were also able to identify where bottlenecks in software, services and architecture exist
- **Technology proponents** were able to do “at-scale” testing to inform software and service evolution

The most recent challenge was a US **capacity** mini-challenge: presentation

What is a “mini-challenge”?

For ATLAS, we found great benefit in the pre-DC24 testing we undertook and realized that having easy to use tools to run “mini-challenges” on demand would be very powerful.

What do we mean by “mini-challenge”. Here is a possible definition:

- A mini-challenge is a **lightweight** way to test capacity or capability with one or more sites utilizing their production systems.

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.

- What is critical is that this should NOT require any expert involvement which currently prevents on-demand testing being easy to do.

Using Challenges and Mini-Challenges to Drive Improvements

We are unsure exactly which technologies, architectures and infrastructures will best meet our HL-LHC scientific needs in an affordable way.

We are using the WLCG Data Challenges as a way to evaluate our progress at scale, using production resources, identifying beneficial capabilities that can help us reach our goals.

However, the 2-3 year cadence for the WLCG Data Challenges is **too long** to wait to evaluate capabilities, technologies, hardware and software and thus we are emphasizing mini-challenges for quicker feedback.

To execute mini-challenges, we need capability/technology advocates to be able to understand the cost-benefit quickly and allow time for improvements to get into our production systems for testing in the next Data Challenge.

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)

Goals for DC27

For DC27, we have some milestones 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)
- At last one example of exploiting ESnet's HighTouch data
- Incorporate Tape as part of the challenge (tape was NOT part of DC24)

All of these areas could benefit from regular testing via mini-challenges

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 ~5 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.

- **Five 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.

Status & Plans for Scitags

The Scitags Initiative has been underway for a while with a goal of allowing traffic identification anywhere in the network.

- We have an IETF draft showing what we are doing

https://scitags.org

DC27 target is to have 80% of our traffic identified by SciTags (fireflies or packets)

- Fireflies are UDP JSON-formatted packets targeting destination port 10514 which contain the experiment (owner) and activity as well as the associated flow information (`src-ip,src-port,dst-ip,dst-port, protocol`)
- We also have targeted IPv6 packets to contain the same information in the "flow-label" field and next steps will be getting that production ready
- We are working on packet marking at 400Gbps for SC25 coming in November
- **dCache** is our next targeted storage infrastructure to enable (v11.2.0)

Testing, Benchmarking and Documenting Tunings

Many of our sites have tuned and documented site software and services but these may have last been tested 5, 10 or even more years ago.

Give the drastic changes in operating systems and software over even a few years, it is important to revisit this area.

We are seeking **sites** and **experiment** and **application teams** to benchmark & optimize various tunings for many different components in use: widely used applications, operating system, storage, computing and networking.

The **goal** is to create a new, **current** reference set of tunings available (perhaps added to the ESnet Fasterdata pages; maybe augment TuneD).

So far, we have advocates from ESnet and some of the USATLAS sites interested…all with an interest are welcome to join.

Traffic Pacing

One way to help address the challenge for HEP storage endpoints to utilize the network efficiently and fully is traffic (packet) pacing.

- Traffic pacing means sending packets at a specific rate, corresponding to to some fraction of the total network bandwidth.
- Without traffic pacing, network packets are emitted by the network interface in **bursts**, corresponding to the wire speed of the interface.
  - **Problem**: microbursts of packets can cause buffer overflows
  - The impact on TCP throughput, especially for high-bandwidth transfers on long network paths can be **significant**.

How? Traffic pacing can be simply enabled and controlled by the Linux 'tc' application, part of the 'iproute' package or TCP congestion protocols like BBR include this already.

The challenge is not in enabling the pacing so much as determining what the pacing should be for a given host and transfer…

Jumbo Frames for Improved Transfer Bandwidth

Using jumbo frames might help increase our usable bandwidth for remote transfers.

ESnet has done testing a while ago showing advantages:

Single stream

- Jumbo frames are 3x faster on 100G hosts
- Jumbo frames are about 15% faster on 10G hosts

8 streams:

- Jumbo frames are about 25% faster on 100G hosts
- Jumbo frames are the same as 1500B 10G hosts

However, sites and even some networks have been hesitant to enable jumbo frames because of associated problems there were seen when many networks and applications were not supporting it well.

CERN EOS Jumbo Frame Testing with CMS

CMS Testing plan with Jumbo frames January 2025

Request: WLCG monitoring with white background

CMS Rucio: EOSPILOT –> FNAL

Data per test:

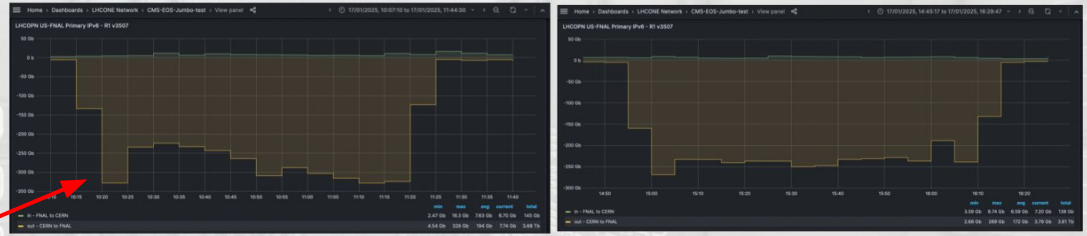
± 126TB reads

- **FTS settings 2000/2000 min/ max**

Key Observations:

- **20-25% faster transfers** with Jumbo Frames for Rucio Activity.
- **Significant performance boost**, especially over longer distances.

Standard Frames — MTU 1500



300 Gb

Jumbo Frames — MTU 9000

400 Gb

Conclusion: Jumbo Frames improve **long-distance** transfers (Geneva-Chicago, RTT 130ms) with **20-25% more bandwidth.**

Maria Arsuaga-Rios - WLCG Doma 19/02/2025

11

Jumbo Frames for Improved Transfer Bandwidth

USATLAS was planning our own tests when we hear about the CERN CMS tests

Our colleagues at CERN were interested in testing ATLAS jumbo frames on long distance paths and this seemed to be a better test for USATLAS to participate in.

The overall results of the testing involving CERN are shown in the WLCG DOMA slides from Maria Arsuaga-Rios (February 2025).

ATLAS-RUCIO: CERN to NET2 and BNL transfers via RUCIO.

- Both NET2 and BNL already had JUMBO frames.
- Transfer 104 TB from CERN to NET2 and from CERN to BNL to evaluate performance with 3-4GB file sizes with and without jumbo frames (CERN end)

Results summary (details in Google doc):

- **Jumbo frames didn't adversely affect site performance or operations**
- **NET2**: a **12% throughput improvement** with Jumbo, **BNL**: no change in performance
- BNL has other bottlenecks than the network (unusual storage/net topology)
- Needs more work and we will likely plan another mini-challenge in 2026 for US sites

Mini-Challenges and Ongoing Testing

As noted before, WLCG plans to have regular mini-challenges going forward.

These will come in two types:

- **Capacity mini-challenges** demonstrate site INPUT/OUTPUT capacity and are used to both benchmark sites and to identify bottlenecks.
- **Capability mini-challenges** demonstrate feasibility and benefits for new technologies, infrastructures and new capabilities

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

We continue to solicit advocates and mini-challenge planners/operators.

- See Google folder
- If a capability you are interested in is not there, add it!

In general, we plan to continue both types of mini-challenges (capability and capacity) through the last WLCG Data challenge in 2029.

Update: USATLAS Capacity Challenges in 2024 and 2025

Comparing: Previous vs Current **capacity** mini-challenge

Site	Fall 2024 (Dec)	Summer 2025 (Aug)	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**

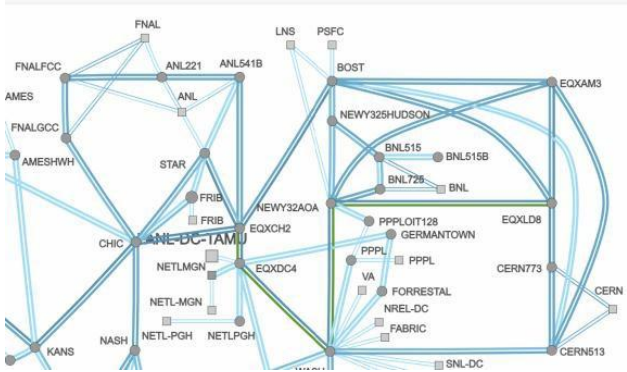
Observation from Joint USATLAS/USCMS Capacity Test

Test: From BNL and NET2 to AGLT2, MWT2 and SWT2 (UTA) and FNAL to all USCMS Tier-2s August 27, 2025

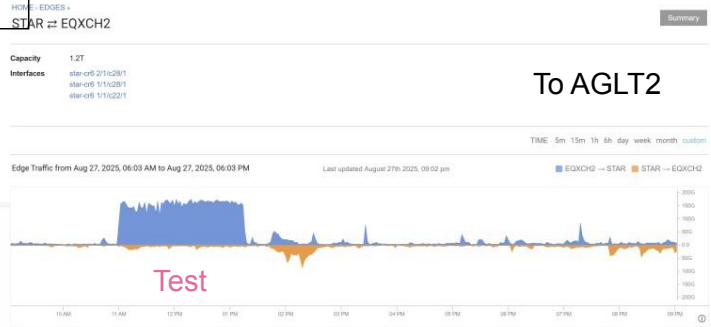
No obvious hotspots have been observed. The destination sites were obtaining data at their expected site rates.

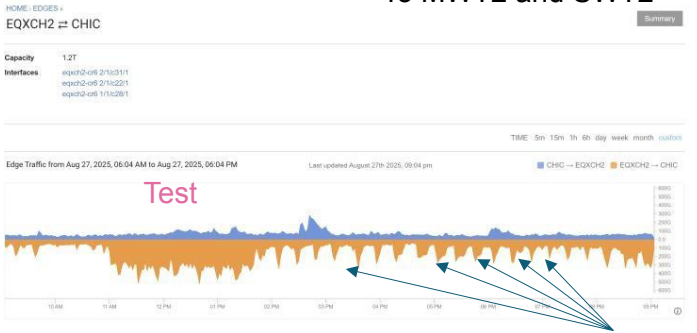
Periodic large data from CERN to FNAL is observed during test test although it didn't affect the data rate for USATLAS sites.

https://my.es.net/



ESNET monitors





Curious periodic traffic?

Near To Mid-Term Plans

In general, continued mini-challenges for both capacity and capability

- USATLAS wants to find the new “maximum” between its Tier-1 at BNL and the various Tier-2s and track site evolution (regression?)
- We are starting to ramp-up network orchestration work (See Harvey’s GNA-g Keynote) and want to include SENSE/Rucio in future ATLAS and CMS mini-challenges.
- Hoping to see some results for:
  - Host tuning recommendations for storage and clients.
  - Packet-marking at 400 Gbps.
  - SDN related work (Netguard/Wirebird VPN, SENSE/Rucio, NOTED, others).
  - Further jumbo frame testing including to/from Asia, Europe and North/South America.
  - New hardware evaluation for WLCG (NVMe devices, new motherboards/buses, new memory, etc)
  - Tape usage optimization and use-cases

Last Note on Monitoring

One very important aspect of all the various challenges we have been undertaking is the critical role monitoring plays.

If you can't monitor it it…you can't manage, debug or understand it

We need to continually verify and validate our monitoring

- It is often not accurate in what is being measured.
- It can be brittle and components fail or disappear over time.
- There may not be existing monitoring for critical items necessary to understand the systems be evaluated.
- New capabilities often require new associated monitoring to be developed.

Our Data Challenges and mini-challenges have improved our monitoring a LOT!

Summary

We need to **clarify** and **expand** existing plans, mini-challenges and goals for the next year and for DC27

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

This work I have described can benefit from and contribute to the larger community, e.g., for a “Global Research Platform”

Question, Comments, Discussion?

Acknowledgements

Thanks to Diego Davila, Hiro Ito and Maria Arsuaga-Rios for their contributions

We would like to thank the **WLCG**, **HEPiX**, **perfSONAR, USLHC** 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 Slides