WLCG DOMA
News and status

Mario.Lassnig@cern.ch
Christoph.Wissing@desy.de
for the DOMA community

GDB
2023-07-12
https://indico.cern.ch/event/1225114/

DOMA - Main Themes

Preparation of upcoming data challenge DC24

High level discussions in the DOMA general meeting

Accompanying technical development

Token based authentication for data transfers
REST API for archival storage
SDN technology for WAN transfers (NOTED, SENSE, ALTO …)
Network packet and flow marking

Mostly discussed in the DOMA BDT Bulk Data Transfer meetings

Main meeting slot & Indico

Meetings are typically Wednesday at 16:00 or 16:30 (CERN time)
DOMA General typically scheduled for last Wednesday of the month
Indico category https://indico.cern.ch/category/10360/

Data challenges for HL-LHC

DOMA mandated to execute the challenges

With increasing capacity & complexity

WLCG objectives

Planning document

Export of RAW data from CERN to the T1s

Data processing

Roughly **bi-annual steps** until HL-LHC

Accompanying R&D programme

2020 estimation of HL-LHC needs

4.8 Tbps of total network capacity (minimal)

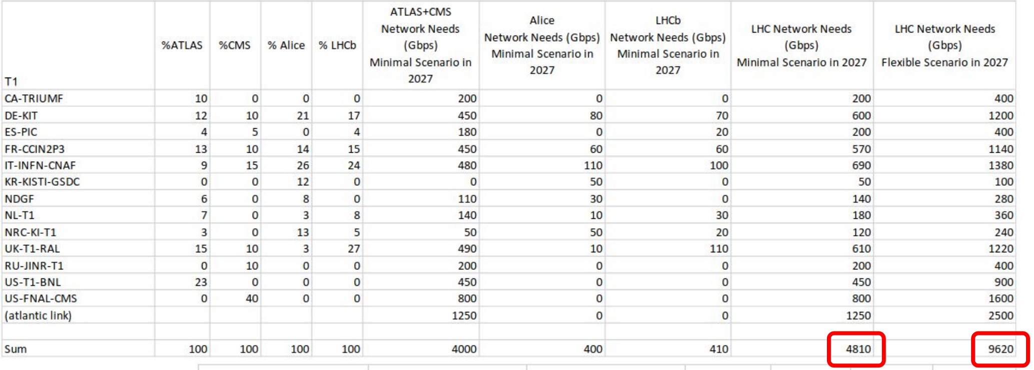
ATLAS & CMS 400 Gbps flat

ALICE & LHCb 100 Gbps flat

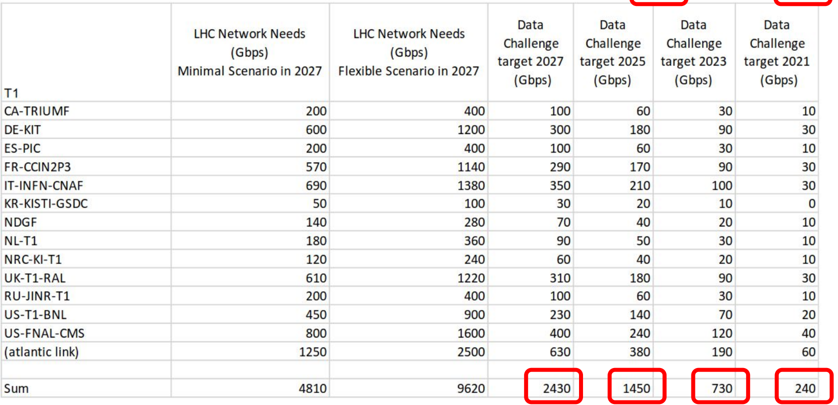
x2 to absorb expected bursts

x2 overprovisioning

Flexible model adds another factor of 2



T1	%ATLAS	%CMS	% Alice	% LHCb	ATLAS+CMS Network Needs (Gbps) Minimal Scenario in 2027	Alice Network Needs (Gbps) Minimal Scenario in 2027	LHCb Network Needs (Gbps) Minimal Scenario in 2027	LHC Network Needs (Gbps) Minimal Scenario in 2027	LHC Network Needs (Gbps) Flexible Scenario in 2027
CA-TRIUMF	10	0	0	0	200	0	0	200	400
DE-KIT	12	10	21	17	450	80	70	600	1200
ES-PIC	4	5	0	4	180	0	20	200	400
FR-CCIN2P3	13	10	14	15	450	60	60	570	1140
IT-INFN-CNAF	9	15	26	24	480	110	100	690	1380
KR-KISTI-GSDC	0	0	12	0	0	50	0	50	100
NDGF	6	0	8	0	110	30	0	140	280
NL-T1	7	0	3	8	140	10	30	180	360
NRC-KI-T1	3	0	13	5	50	50	20	120	240
UK-T1-RAL	15	10	3	27	490	10	110	610	1220
RU-JINR-T1	0	10	0	0	200	0	0	200	400
US-T1-BNL	23	0	0	0	450	0	0	450	900
US-FNAL-CMS	0	40	0	0	800	0	0	800	1600
(atlantic link)					1250	0	0	1250	2500
Sum	100	100	100	100	4000	400	410	4810	9620



T1	LHC Network Needs (Gbps) Minimal Scenario in 2027	LHC Network Needs (Gbps) Flexible Scenario in 2027	Data Challenge target 2027 (Gbps)	Data Challenge target 2025 (Gbps)	Data Challenge target 2023 (Gbps)	Data Challenge target 2021 (Gbps)
CA-TRIUMF	200	400	100	60	30	10
DE-KIT	600	1200	300	180	90	30
ES-PIC	200	400	100	60	30	10
FR-CCIN2P3	570	1140	290	170	90	30
IT-INFN-CNAF	690	1380	350	210	100	30
KR-KISTI-GSDC	50	100	30	20	10	0
NDGF	140	280	70	40	20	10
NL-T1	180	360	90	50	30	10
NRC-KI-T1	120	240	60	40	20	10
UK-T1-RAL	610	1220	310	180	90	30
RU-JINR-T1	200	400	100	60	30	10
US-T1-BNL	450	900	230	140	70	20
US-FNAL-CMS	800	1600	400	240	120	40
(atlantic link)	1250	2500	630	380	190	60
Sum	4810	9620	2430	1450	730	240

Ingress-only rate

DC21 Data rate table

ATLAS & CMS T0 to T1 per experiment

350PB RAW, taken and distributed during typical LHC uptime of 7M seconds / 3 months (50GB/s aka. 400Gbps)
Another 100Gb/s estimated for prompt reconstruction data (AOD, other derived output)
In total approximately 1Tbps for CMS and ATLAS together

ALICE & LHCb

100 Gbps per experiment estimated from Run-3 rates

Minimal model

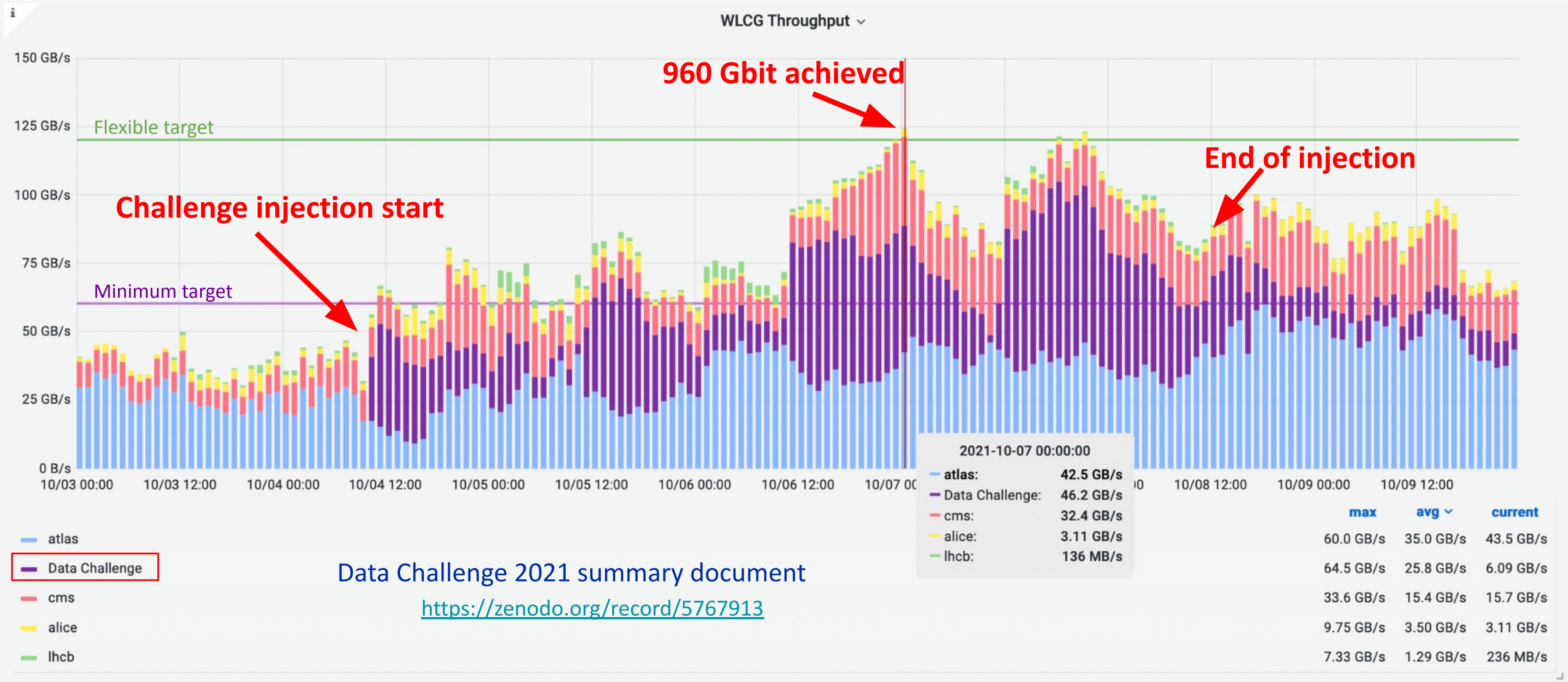
∑ (ATLAS,ALICE,CMS,LHCb) *2 (for bursts) *2 (overprovisioning) = **4.8Tbps**

Flexible model

Assumes reading of data from above for reprocessing/reconstruction within 3 months
Means doubling the Minimal Model: **9.6Tbps**
However data flows from the T1s to T2s and T1s!

No MC production flows nor re-creation of derived data in the 2021 modelling!

DC21 goal: 10% of HL-LHC



960 Gbit achieved

Challenge injection start

End of injection

2021-10-07 00:00:00

atlas:	42.5 GB/s
Data Challenge:	46.2 GB/s
cms:	32.4 GB/s
alice:	3.11 GB/s
lhcb:	136 MB/s

	max	avg	current
atlas	60.0 GB/s	35.0 GB/s	43.5 GB/s
Data Challenge	64.5 GB/s	25.8 GB/s	6.09 GB/s
cms	33.6 GB/s	15.4 GB/s	15.7 GB/s
alice	9.75 GB/s	3.50 GB/s	3.11 GB/s
lhcb	7.33 GB/s	1.29 GB/s	236 MB/s

Data Challenge 2021 summary document

https://zenodo.org/record/5767913

Data rate complexity

Data rate experience from DC21

Higher complexity of data flows than assumed has become evident

Include feedback from the experiments and the network community

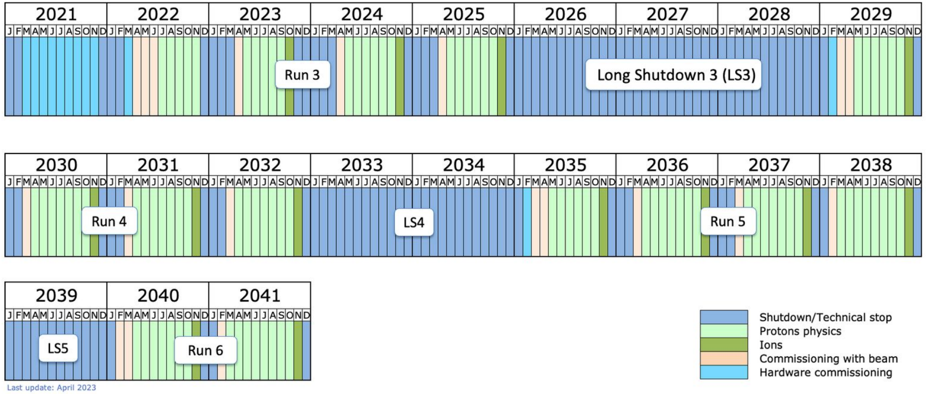
Mixing of ingress/egress values was very confusing

More complex setup has three major data flows

RAW export, prompt reconstruction/derivation export …	Tier-0 to Tier-1	Unidirectional
Reconstruction, Reprocessing, Simulation, Derivations, …	Tier-1+2 to Tier-1+2	Bi-directional
Data consolidation, recovery operations, …	Tier-1+2 to Tier-1+2	Bi-directional

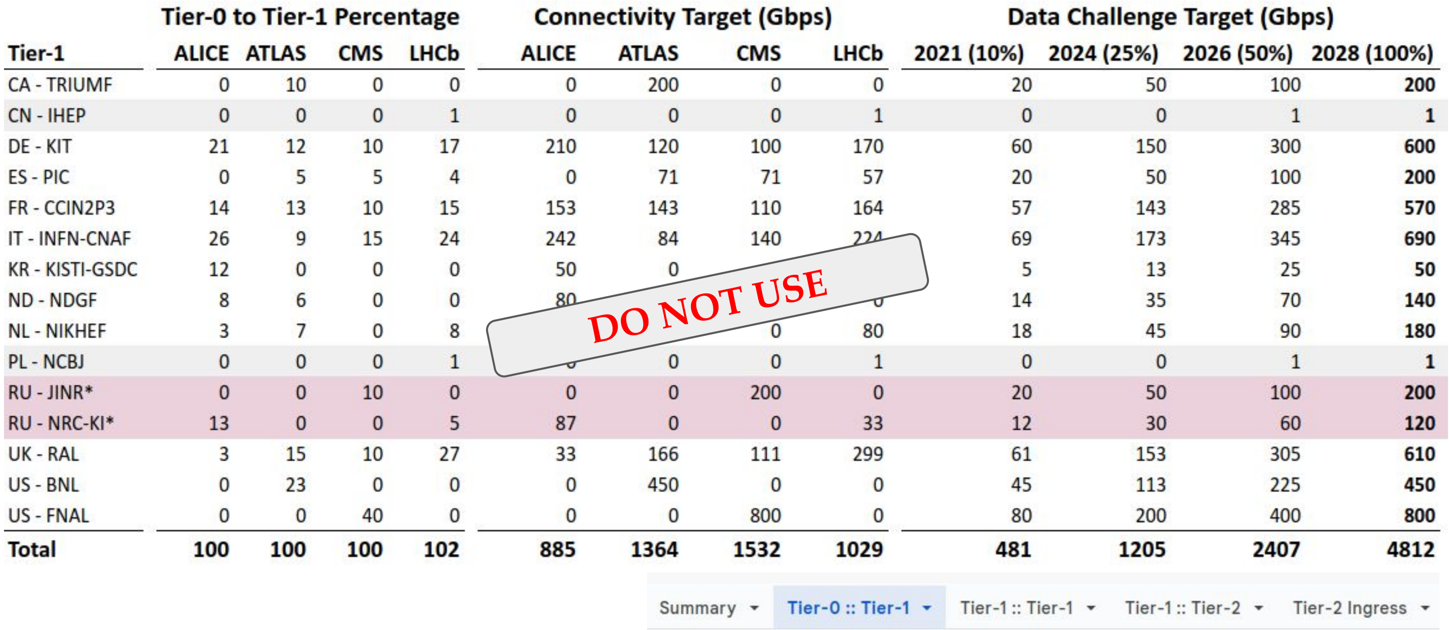
Assume the following steps

2021 → 10%
2024 → 25%
2026 → 50%
2028 → 100%



Last update: April 2023

Example table / WIP



	Tier-0 to Tier-1 Percentage				Connectivity Target (Gbps)				Data Challenge Target (Gbps)			
Tier-1	ALICE	ATLAS	CMS	LHCb	ALICE	ATLAS	CMS	LHCb	2021 (10%)	2024 (25%)	2026 (50%)	2028 (100%)
CA - TRIUMF	0	10	0	0	0	200	0	0	20	50	100	**200**
CN - IHEP	0	0	0	1	0	0	0	1	0	0	1	**1**
DE - KIT	21	12	10	17	210	120	100	170	60	150	300	**600**
ES - PIC	0	5	5	4	0	71	71	57	20	50	100	**200**
FR - CCIN2P3	14	13	10	15	153	143	110	164	57	143	285	**570**
IT - INFN-CNAF	26	9	15	24	242	84	140	224	69	173	345	**690**
KR - KISTI-GSDC	12	0	0	0	50	0	[illegible]	[illegible]	5	13	25	**50**
ND - NDGF	8	6	0	0	80	[illegible]	[illegible]	0	14	35	70	**140**
NL - NIKHEF	3	7	0	8	[illegible]	[illegible]	0	80	18	45	90	**180**
PL - NCBJ	0	0	0	1	0	0	0	1	0	0	1	**1**
RU - JINR*	0	0	10	0	0	0	200	0	20	50	100	**200**
RU - NRC-KI*	13	0	0	5	87	0	0	33	12	30	60	**120**
UK - RAL	3	15	10	27	33	166	111	299	61	153	305	**610**
US - BNL	0	23	0	0	0	450	0	0	45	113	225	**450**
US - FNAL	0	0	40	0	0	0	800	0	80	200	400	**800**
Total	**100**	**100**	**100**	**102**	**885**	**1364**	**1532**	**1029**	**481**	**1205**	**2407**	**4812**

DO NOT USE

Straightforward application of the original "minimal model" for the Tier-0 :: Tier-1 case

Already includes the two new Tier-1s (CN - IHEP, PL - NCBJ) with fake numbers

Still includes the Russian Tier-1s (RU - JINR, RU - NRC-KI) with the 2020 numbers

LHC Experiment Questionnaire - Q3 2022

When would you like to do DC23-DC24?

After the processing of the last 2023 **Heavy Ion run** has finished
Before 2024 **pp run** starts ⇨ early 2024
Not during ISGC week (typically late March)

What would you like to do?

Specific focus on the test of **SE tokens** for storage, and **migration to IAM**
Monitoring with **IPv6 flow labels**
Demonstrate **SDNs** (*SENSE, AutoGOLE, NOTED, ALTO/TCN*) on selected production sites
Tape challenges were part of the Run-3 export commissioning, necessary to repeat?
Test **peering** with commercial clouds if possible

How do you want to do it?

Start with a series of distributed, constrained, and isolated **ramp-up challenges**
- Independently organised, and report via WLCG/DOMA

Hardware purchasing greatly affects data challenge scope, influence on sites non-negligible
- **Early integration** of Tier-1s and Tier-2s in the planning
- Instead of a Data Challenge, possibility of **stress tests** instead?
- **Revisit the original requirements**, reduce to 20% or 25% challenge?

Kindle discussion with **non-LHC experiments** for possible future combined Data Challenges

DC24 Planning Document

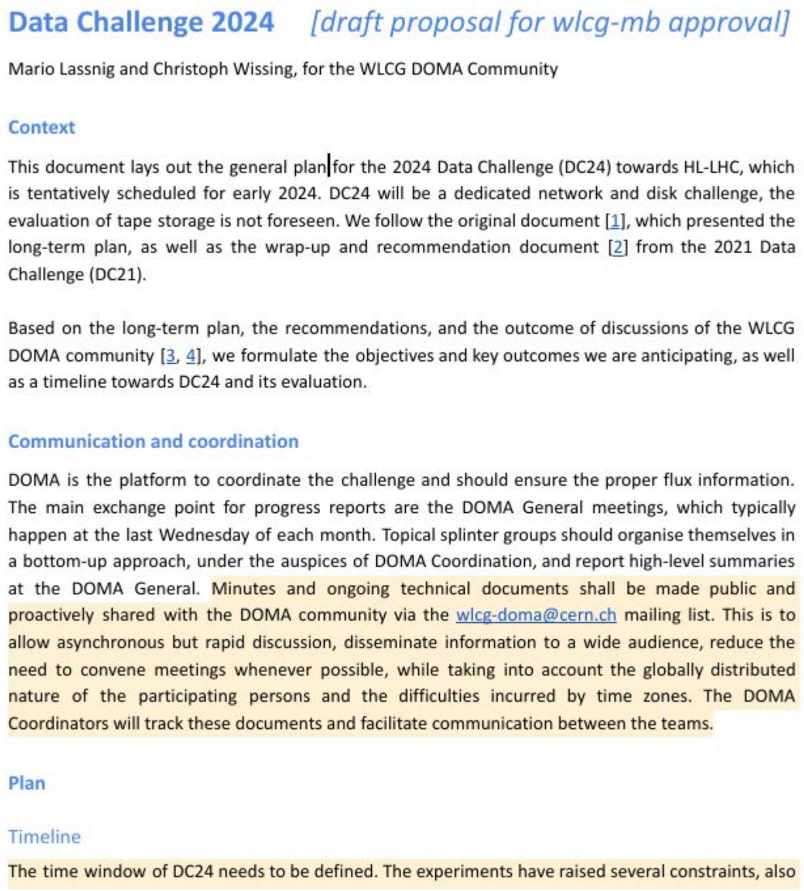
Discussed during DOMA General meeting on Feb 1st

Presented in the WLCG MB on Feb 14th

- Target rate should be **25%**
- Main request: New table with updated numbers

Final version still dependent on new data rate table

- Expected by end of August
- Experiment summary documents



Data Challenge 2024 *[draft proposal for wlcg-mb approval]*

Mario Lassnig and Christoph Wissing, for the WLCG DOMA Community

Context

This document lays out the general plan for the 2024 Data Challenge (DC24) towards HL-LHC, which is tentatively scheduled for early 2024. DC24 will be a dedicated network and disk challenge, the evaluation of tape storage is not foreseen. We follow the original document [1], which presented the long-term plan, as well as the wrap-up and recommendation document [2] from the 2021 Data Challenge (DC21).

Based on the long-term plan, the recommendations, and the outcome of discussions of the WLCG DOMA community [3, 4], we formulate the objectives and key outcomes we are anticipating, as well as a timeline towards DC24 and its evaluation.

Communication and coordination

DOMA is the platform to coordinate the challenge and should ensure the proper flux information. The main exchange point for progress reports are the DOMA General meetings, which typically happen at the last Wednesday of each month. Topical splinter groups should organise themselves in a bottom-up approach, under the auspices of DOMA Coordination, and report high-level summaries at the DOMA General. Minutes and ongoing technical documents shall be made public and proactively shared with the DOMA community via the wlcg-doma@cern.ch mailing list. This is to allow asynchronous but rapid discussion, disseminate information to a wide audience, reduce the need to convene meetings whenever possible, while taking into account the globally distributed nature of the participating persons and the difficulties incurred by time zones. The DOMA Coordinators will track these documents and facilitate communication between the teams.

Plan

Timeline

The time window of DC24 needs to be defined. The experiments have raised several constraints, also

Non-LHC Participation in DC24

Interest in the wider HEP community to join DC24

Belle II, DUNE, JUNO

Perhaps SKA (radio astronomy)

Involved sites are often supporting also LHC experiments

Overall traffic from non-LHC expected to be small compared to LHC

Parts of the traffic going through LHCONE networks

Direction often in the opposite direction, e.g.

- LHC RAW data: From Europe to US and Asia
- DUNE: From US to Europe (and Asia)
- Belle II & Juno: From Asia to Europe and US

Monitoring

Good common(!) monitoring of LHC traffic already challenging

Common dashboard with non-LHC experiments would be great, but quite some effort

However (low level) monitoring of network providers should show these activities

BDT topics

Token based authentication for data transfers

- Decide about porting features of GsiFTP to Http/WebDAV (e.g. multi-stream) if necessary
- Coordinate timeline with WLCG AuthZ working group

Tape REST API

- Roll out plan for all T1s

WLCG data transfer monitoring

- Focus Xrootd monitoring deployment initially at CERN and FNAL (main sources for CMS pileup mixing)

Collaboration beyond LHC experiments

- A number of topics have been addressed in the context of ESCAPE
  - Joining efforts where same tools are being used (e.g. Rucio, FTS, Dirac …)
  - Analysis facilities
  - Usage of shared sites & infrastructures, e.g. storage consolidation
  - Common AAI solutions
- Foster exchange with "close" projects, Belle-2, DUNE, SKA

HTTP REST API for Archival Storage

A REST API should replace SRM at tape (archival) storage endpoints

- Phase out last SRM use case
- Provide a simpler interface

Reference documentation:

- Worked out by all major WLCG storage middleware providers: EOSCTA, dCache, StoRM, Gfal2 and FTS

Deployment status

- ATLAS: Ongoing campaign, CERN/RAL/KIT in production
- CMS: Campaign just started

Tape exercises are optional in DC24

- Experiments are free to include tests

Flow Labeling & Packet Marking

Identify certain traffic, e.g. by experiment or major transfer activity

Flow Labeling via UDP Fireflies and Packet Marking

- Fireflies are UDP packets in Syslog format with a defined schema
- Packets to be sent to regional or global collectors for monitoring

Packet marking intended to use flow label field

- Only available in IPv6
- Enables tracking of packets by 'owner' or experiment

DC24 is a good opportunity to test things on a larger scale in production-like environment

Transfers with token-based authentication

Token support needed in all parts of involved middleware

- Storage systems (dCache, ECHO, EOS, StoRM, XrootD…)
- FTS
- Rucio (ATLAS and CMS), Dirac (LHCb)

Most likely Run-3 scenario

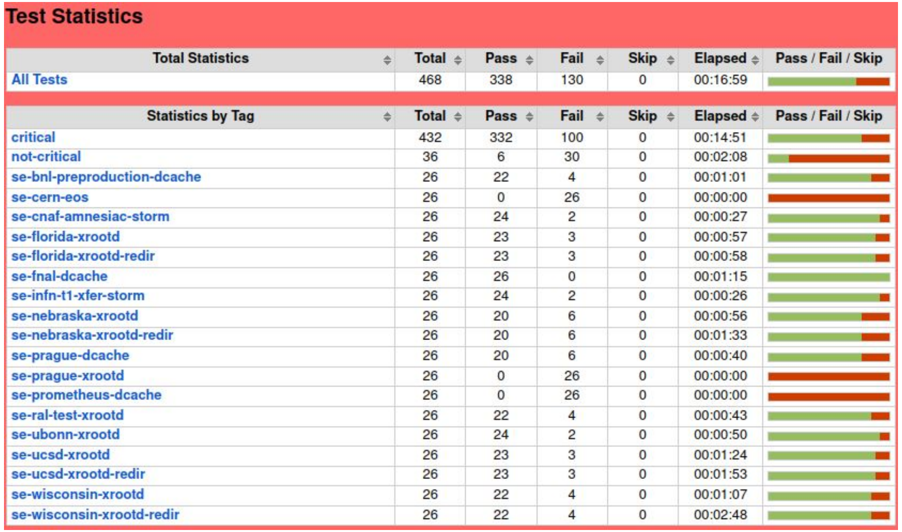
- Dual mode, supporting X509 and Tokens in parallel
- Move more and more transfers to Tokens

Roll-out status

- ATLAS & CMS working with early adopters

Tokens planned for all dedicated DC24 injections

- Scale testing of token infrastructure



Test Statistics

Total Statistics	Total	Pass	Fail	Skip	Elapsed	Pass / Fail / Skip
All Tests	468	338	130	0	00:16:59	🟩🟥

Statistics by Tag	Total	Pass	Fail	Skip	Elapsed	Pass / Fail / Skip
critical	432	332	100	0	00:14:51	🟩🟥
not-critical	36	6	30	0	00:02:08	🟩🟥
se-bnl-preproduction-dcache	26	22	4	0	00:01:01	🟩🟥
se-cern-eos	26	0	26	0	00:00:00	🟥
se-cnaf-amnesiac-storm	26	24	2	0	00:00:27	🟩🟥
se-florida-xrootd	26	23	3	0	00:00:57	🟩🟥
se-florida-xrootd-redir	26	23	3	0	00:00:58	🟩🟥
se-fnal-dcache	26	26	0	0	00:01:15	🟩
se-infn-t1-xfer-storm	26	24	2	0	00:00:26	🟩🟥
se-nebraska-xrootd	26	20	6	0	00:00:56	🟩🟥
se-nebraska-xrootd-redir	26	20	6	0	00:01:33	🟩🟥
se-prague-dcache	26	20	6	0	00:00:40	🟩🟥
se-prague-xrootd	26	0	26	0	00:00:00	🟥
se-prometheus-dcache	26	0	26	0	00:00:00	🟥
se-ral-test-xrootd	26	22	4	0	00:00:43	🟩🟥
se-ubonn-xrootd	26	24	2	0	00:00:50	🟩🟥
se-ucsd-xrootd	26	23	3	0	00:01:24	🟩🟥
se-ucsd-xrootd-redir	26	23	3	0	00:01:53	🟩🟥
se-wisconsin-xrootd	26	22	4	0	00:01:07	🟩🟥
se-wisconsin-xrootd-redir	26	22	4	0	00:02:48	🟩🟥

Network R&D

NOTED

- Monitor link saturation and predict the behaviour of the applications
- When NOTED detects that the link is going to be congested provides a dynamic circuit using AutoGOLE/SENSE
- Ongoing work in decision making, improving the forecasts, monitoring integration, FTS integration

AutoGOLE/SENSE

- End-to-end service to dynamically procure VPNs between routers to enforce a given path
- Implement network QoS to prioritise transfers at the router level
- Ongoing integration work with Rucio

ALTO/TCN

- Application-Layer Traffic Optimization provides means to to obtain network information
- Exploit this network information in higher-level long-term schedules (FTS / Rucio)

Jumbo frames

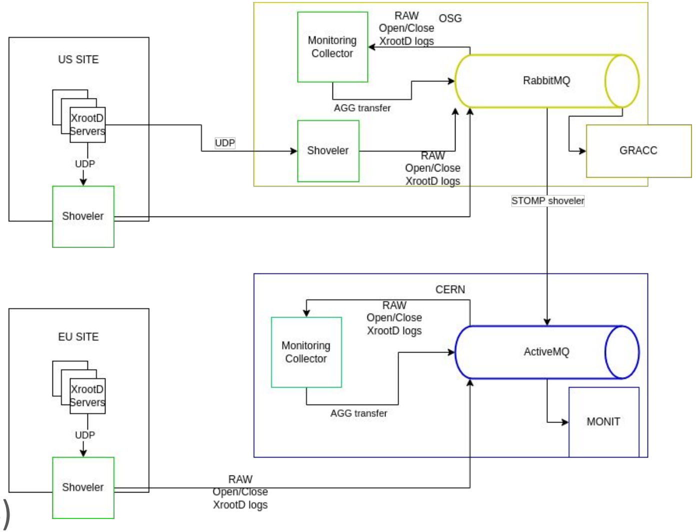
- CERN is working on setting up Jumbo frame evaluation for a set of EOS servers

and many more (DTNs, caches, …)

WLCG monitoring

Improvement of XRootD traffic monitoring

Redesign of present XRootD implementation
New components: XRootD shoveler & collector
Integration with dCache XRootD stack
Integration of ALICE XRootD monitoring flow



Site aggregated network monitoring

Identified as missing piece in DC21
Needs collection of data on site level
WLCG deployment campaign (needs some push to conclude by DC24)
- Recipes for sites
- Data published by a site webservice
- Configuration announced via CRIC

The most important slide

The date for DC24 has been fixed

- Two weeks: **February 26 (Monday) to March 8 (Friday) 2024**
- Only subject to change upon **major** interfering events by the experiments

DC24 Preparation Workshop

- Two days: **November 9 (Thursday) to November 10 (Friday) 2023**
- In the same week: pre-GDB on tapes (Tue) and GDB (Wed)
- Hybrid event, in-person at CERN
  - Compose near-to-complete schedule for DC24
  - Identify remaining set of issues
  - Prioritize the work on open items
- Updates on DC24 operation by the experiments
- Results from the various ramp-up challenges
  - Network and storage infrastructure
  - Readiness regarding token based authentication (Rucio, FTS, storage, …)
  - Status of the SDNs (SENSE, ALTO, NOTED, …)
  - Monitoring capabilities (MONIT, Fireflies, …)
  - Feedback & plans by sites

Extras

Contact persons

Based on previous communication, we have these contact persons for the confirmed participants

CMS	Katy Ellis
ATLAS	Alessandra Forti
LHCb	Christophe Haen
ALICE	Latchezar Betev
Belle II	Silvio Pardi
DUNE	Doug Benjamin
JUNO	Xiaomei Zhang

Please let us know ASAP if this list needs modification!

We have a good set of DC24 "observer" communities, we welcome all feedback and suggestions!