Data Challenge 2024 – CMS Activities

Christoph Wissing[1,*], *Rahul* Chauhan[2], *Katy* Ellis[3], *Andres Manrique* Ardila[4], *Hasan* Ozturk[2], *Panos* Paparrigopolos[2], and *Garyfallia* Paspalaki[5]
for the CMS Collaboration

[1]Deutsches Elektronen-Synchrotron DESY, Notkestr. 85, 22603 Hamburg, Germany
[2]CERN, Meyrin, Switzerland
[3]Rutherford Appleton Laboratory, Harwell Campus, United Kingdom
[4]University of Wisconsin - Madison, 1150 University Avenue, Madison, WI 53706, United States of America
[5]Purdue University, 1396 Physics Building West Lafayette, IN 47907-139, United States of America

Abstract.
To verify the readiness of the data distribution infrastructure for the HL-LHC, which is planned to start in 2030, WLCG is organizing a series of data challenges with increasing throughput and complexity. This presentation addresses the contribution of CMS to Data Challenge 2024, which aims to reach 25% of the expected network throughput of the HL-LHC. During the challenge CMS tested various network flows, from the RAW data distribution to the "flexible" model, which adds network traffic resulting from data reprocessing and MC production between most CMS sites.
The overall throughput targets were met on the global scale utilizing several hundred links. Valuable information was gathered regarding scaling capabilities of key central services such as Rucio and FTS. During the challenge about half of the transferred volume was carried out via token based authentication. In general sufficient performance of individual links was observed and sites coped with the target throughput. For links that did not reach the target, attempts were made to identify the bottleneck, whether in the transfer tools, the network link, the involved storage systems or other component.

1 Introduction

The CMS experiment [1, 2] is a multi-purpose detector that is installed at the Large Hadron Collider (LHC), presently the most powerful proton-proton collider in the world. CMS studies proton-proton scattering events at a center of mass energy of $\sqrt{s}$=13.6 TeV in order to precisely measure parameters of the Standard Model and constrain or discover processes beyond the Standard Model. The next major phase of the LHC is the High Luminosity LHC (HL-LHC), which is presently scheduled to start in 2030. Due to the stronger focusing of the proton beams the scattering events will grow significantly in size because of the increased number of tracks. Since also the rate, at which the collisions are permanently stored for later analysis, is expected to grow as well, the demand for disk and archival storage as well as network bandwidth to transfer the data will increase roughly by an order of magnitude.

*e-mail: christoph.wissing@desy.de

2 Worldwide LHC Computing Grid

The data produced by the experiments at the LHC are stored and processed on a globally distributed infrastructure, the Worldwide LHC Computing Grid (WLCG) [3]. The WLCG is organized in a tiered structure with the Tier-0 being at CERN, where the LHC experiments, ALICE, ATLAS, CMS and LHCb, are located. CMS is presently supported by six Tier-1 sites: KIT in Germany, PIC in Spain, CCIN2P3 in France, CNAF in Italy, RAL in the UK and FNAL in the US. For political reasons the site at JINR in Russia cannot be used as a Tier-1, but is used opportunistically without storing any unique data. Beyond substantial CPU and disk storage Tier-1 sites provide archival storage and support on a 24-7 basis. Tier-2 sites provide CPU and disk resources and commit to business hours support. CMS is supported at around 45 Tier-2 sites.

WLCG is a collaboration between the resource providers, the middleware providers and the experiments. Since participating institutions and computing centers are often involved in particle physics experiments other than the LHC experiments, the collaboration has been extended in the recent past. The Belle II experiment, that operates at KEK in Japan, as well as the DUNE experiment, which will be constructed at FNAL and SURF in the USA, have become stakeholders in WLCG.

3 WLCG Data Challenge Series

Since the demands for network throughput are expected to rise by roughly an order of magnitude WLCG has mandated a series of data challenges (DC) to demonstrate the readiness of the infrastructure prior to the start of the HL-LHC data taking. The challenges will increase stepwise towards the final capacity and also in terms of storage technology, middleware and monitoring. To date, two challenges have been executed. The first DC of this series was "DC21" in 2021 and targeted 10% of the HL-LHC throughput. A major aspect was the usage of WebDAV as the main WAN transfer protocol that has replaced the legacy GridFTP protocol. "DC24" in February 2024 is the subject of this document. The primary goals were to reach 25% of the expected HL-LHC traffic and to enable tokens for authentication (see section 6.3). The third DC is foreseen at the start of 2027 with a target around 50%, and yet another DC prior to HL-LHC to demonstrate 100% readiness for the HL-LHC throughput.

4 Planning and Preparation

The planning of DC24 started more than a year before the actual challenge with regular discussions in the WLCG DOMA (Data Organization, Management and Access) forum. Since the mandate set goals only at a high level, the experiments together with the sites, the network providers and the middleware development teams took charge to shape a more detailed program for the challenge. The experiments were given the freedom to shape the challenge along their own program of work. However, particularly ATLAS and CMS made an effort to align their activities as much as possible. Both experiments agreed not to include dedicated exercises for tape in the challenge, because archives at the sites had not been significantly upgraded regarding performance and therefore no major progress could be expected with respect to DC21, when archives were tested primarily to meet the Run-3 requirements.

4.1 Selection of Tools

CMS decided to run DC24 within the existing production infrastructure and no services or storage areas were set up just for the purpose of the challenge. Instead all additional load

required to meet the target rates was injected on top of ongoing production data transfers. To steer the injection of this extra traffic the dc_inject tool [4] was selected. The tool was developed by ATLAS during DC21. To reach a configurable transfer rate on a link between two storage endpoints dc_inject creates appropriate rules in the central data management system Rucio [5]. Due to the short lifetime of the subscription data gets removed again soon after the transfer in order not to occupy disk space unnecessarily. During the preparation dc_inject was further developed together by the ATLAS and the CMS data management teams.

Several weeks before the actual challenge a series of pre-challenges was executed. These were smaller in scope and focused typically on just one region. During these exercises valuable experience was gained by the operations teams how to use e.g. the dc_inject tools and to become familiar with monitoring pages.

4.2 Rate Estimates

CMS took a pragmatic approach to the estimation of rates expected in Run-4, of which 25% were targeted in DC24. A set of major workflows was identified:

- Tier-0 exports (data transfers from CERN to Tier-1s)
- Tier-1 exports (data transfers from Tier-1s to Tier-2s for reprocessing)
- MC production Input/Output (data transfers relating to Monte Carlo jobs in either direction among Tier-1s and Tier-2s)
- AAA (unscheduled, streamed transfers using the CMS "Any data, anytime, anywhere" Xrootd storage federation [6])

The Tier-0 export rate was well defined and could be modeled from the parameters published in the CMS DAQ TDR [7] following the same approach outlined in the WLCG mandate document [8]. An annual volume of 350 PB of RAW data must be exported in a pseudo-real time of 7 million seconds LHC uptime, plus some headroom for derived data products. To derive the required network bandwidth a doubling factor was then applied to account for 'bursty' behaviour. The DC24 target was 250 Gbps for the Tier-0 export.

In a simplified model the Tier-1 exports move the same data specified in the Tier-0 export and within the same time period, but from Tier-1s to Tier-2s. Hence the DC24 target was also 250 Gbps, again applying a factor 2.

Using information about data transfers during Run-3 the data transfer rate for MC production and miscellaneous activity was estimated as broadly similar to the rate of data – again resulting in a DC24 target at 250 Gbps.

Finally, the target for data moved by the AAA federation was also approximated at 250 Gbps by extrapolating the fraction of AAA transfers as observed during the ongoing Run-3. The content of these data transfers is not well-understood, as detailed monitoring for XRootD [9] which is the underlying protocol for AAA is still a work in progress. However, it is assumed to be a mixture of 'premix' libraries containing background events, which are streamed only from the largest two sites (CERN & FNAL), T0-jobs routed to Tier-1s but reading from CERN, and jobs that run at a site where the data is not stored and is therefore streamed from a remote source.

In Run-3, the first three flows mentioned above use the Rucio Data Management software and the File Transfer Service (FTS) [10] to replicate data. The fourth 'AAA' workflow instead does direct streamed transfers via XRootD. Lacking a tool to generate AAA traffic in a controlled manner in DC24 the AAA transfers were simulated via FTS transfers, with the intention of generating additional pressure on the main source storage systems at CERN and FNAL.

Executing transfer injections for all four mentioned flows the total DC24 target sums to a throughput rate of 1 Tbps, which corresponds to a "flexible target" sketched in the WLCG mandate document. The "minimal target" was half of it. Reaching the minimal target would imply for CMS the simultaneous execution of flows 1 and 2 to reach 500 Gbps.

The overall rates needed to be distributed over individual network links. For the Tier-0 export the load was split following the pledged tape capacity of the Tier-1s at the beginning of Run-3, which corresponds to 40% for FNAL and roughly 10% for each of the European Tier-1 sites. For the other flows it was assumed that the Run-3 distribution over the various links would also be applicable for HL-LHC.

4.3 Schedule

It was observed during DC21 that one week was insufficient, and therefore DC24 lasted two working weeks. Increasing complexity formed the basis of the scheduling for the different CMS exercises, which are shown in Table 1.

Table 1. Schedule for the execution of the various network flows during the two week of DC24.

Date	12 Feb	13 Feb	14 Feb	15 Feb	16 Feb	17 Feb	18 Feb	19 Feb	20 Feb	21 Feb	22 Feb	23 Feb
	T0 exp	T0 exp	T0 exp T1 exp	T1 exp	T1 exp Prod out	T1 exp Prod out	T1 exp Prod out	AAA	T0 exp T1 exp Prod out AAA	T0 exp T1 exp Prod out AAA	T0 exp T1 exp Prod out AAA	T0 exp T1 exp Prod out AAA
Target [Gbs]	250	250	500	250	500	500	500	250	1000	1000	1000	1000
Target [GB/s]	31	31	62	31	62	62	62	31	125	125	125	125

The first 2 days were devoted the T0 export in order to confirm the most important workflow – moving RAW data from CERN to the Tier-1 sites. On day 3, workflows 1 and 2 were run simultaneously, to test the ability of Tier-1 sites to write and read simultaneously, and on day 4 only workflow 2 was performed to observe the write rates at Tier-2 sites. During days 4-6, which included the weekend, both workflows 2 and 3 were maintained. On day 8 the 'AAA' workflow was tested in isolation. From day 9 until day 12 all workflows were executed in attempt to reach the flexible model target.

This schedule was agreed with the other experiments and in particular with ATLAS, against whom it was important to try to stress our common sites simultaneously and also to have a chance to achieve an overall multi-experiment flexible target of 2.4 Tbps [11].

5 Execution of the challenge

Figure 1 shows the overall throughput and highlights the major operational issues spotted during the 12 days of DC24 which had a significant effect on the transfer rate. In chronological order these were:

- Token refresh problem between FTS and the token issuer. This was mitigated on-the-fly by a change in FTS.
- Small files were used in preference to larger file sizes. In its usual configuration the dc_inject tool picks files from larger datasets first. Running with a lot of small files instead the transfer system tends to become inefficient regarding the throughput rate. On this particular occasion the "small-files-first" option was activated unintentionally but was corrected by restarting the dc-inject tool with a proper configuration after the issue had been spotted and understood.
- Insufficient deletions meant that space ran out at sites and new transfers did not occur. Since the available disk-space on the storages is finite, new transfers can only be initiated

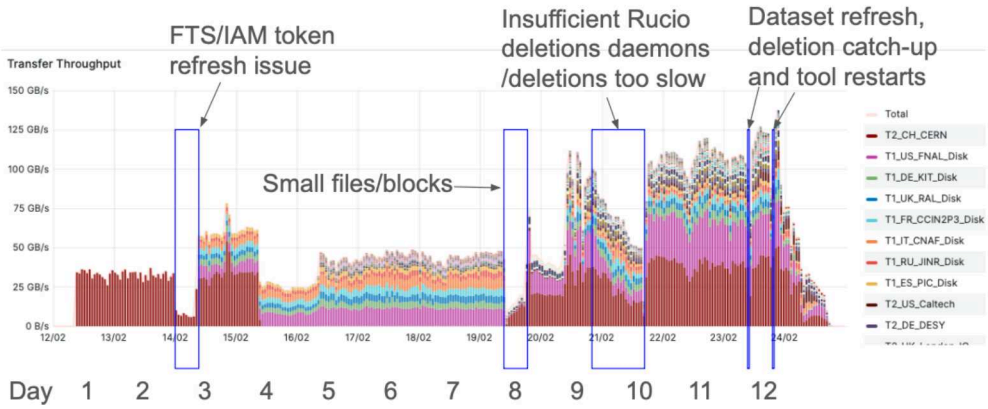

Figure 1. Transfer throughput by source site with some operational issues encountered during DC24.

after deletions and space becomes available. Using Rucio these deletions are orchestrated centrally and are executed remotely. The observed rate degradation could be fixed by increasing the number of Rucio deletion daemons and by targeting dedicated daemons at sites which showed particularly slow deletions.

- Two short periods where the dc-inject tool was stopped and restarted to allow the system to catch up with deletions and refresh the source dataset list.

It should be noted that the highlighted sections with identified operational issues were excluded when calculating average rates achieved by the sites, discussed in the next section.

6 Results

6.1 Overall Throughput

Figure 2 shows the overall CMS network throughput as monitored by FTS compared to the target rates for the given days. The vast majority of the traffic is due to DC24 injections. The target for the first two days was achieved instantaneously. Here the program of pre-challenges paid off, because the links from CERN to the Tier-1s had already been exercised and some FTS parameters were optimized. In addition, the scenario on day three involving both Tier-0 export and Tier-1 export reached the target easily.

The scenario on day 5 targeted a throughput which was achieved already on the third day, but it involved many more network links. This turned out to be more difficult, because FTS had to handle a large number of transfers and also involved links that were not studied during pre-challenges. Therefore the target rate could be met only during peaks. It should be noted that during the weekend (days 6 & 7) the system was kept running mostly unattended and kept going with stable performance.

For day 8, where the original schedule planned for only the AAA test, extra injected traffic between the US Tier-2 sites increased the combined target to 41 GB/s. While the combined goal for day 8 was fulfilled relatively easily it was much more challenging to achieve the goal of the flexible target, running with all scenarios in parallel and extending the number of active network links to about 200. A continuous effort over the last days of the challenge involving tuning of parameters in FTS and Rucio and huge effort by the operation and development teams finally let the throughput surpass the target for the flexible target for several hours.

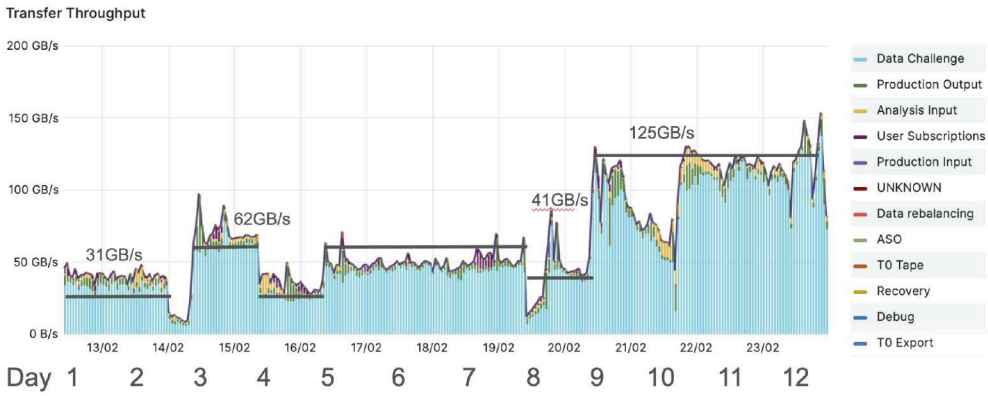

Figure 2. Transfer throughput by activity type in comparison with the target rates.

6.2 Site Performance

For the different scenarios each of the involved network links was expected to contribute to the overall throughput corresponding to the amount of traffic that had been injected with the dc_inject tool. In order to access the performance of the involved sites the expected rates were compared to rates observed by the monitoring infrastructure. Table 2 shows the fraction of the achieved throughput compared with the target for the Tier-1 sites and for each day of the challenge. Despite the fact that JINR could not be treated as an official Tier-1 site for political reasons the site was included in DC24 but limited effort was spent to understand encountered performance degradations. During the first week of the challenge sites reached the expected values or were close to them. A performance degradation for RAL during the first days could be attributed to a cutting of the LHC OPN network link between CERN and RAL. During the final days when all scenarios were active, more sites struggled to hit the targets. However the limitation was typically located in FTS that did not initiate enough transfers. During the challenge the situation was improved for many links by increasing the number of transfers for a link or a site allowed in the FTS configuration. Increasing these limits too much led to a decrease of the overall performance of FTS because the internal components had difficulties to digest the huge amount of requests in the system. On the other hand, a too high number of transfers can lead to an overload of the storage itself as observed at CNAF. For the observed links the provisioned capacity was not exhausted at every site. In fact, towards the end of the challenge several Tier-2 sites explicitly asked to increase the load on their WAN connection to approach nominal available bandwidth. The findings for sites and possible solutions or mitigation strategies are great subjects for so-called "Mini Challenges", that are already agreed to be performed throughout the coming months before the next major Data Challenge. These Mini Challenges are targeted exercises addressing impact of changing of selected configuration parameters or transfers between a small number of sites e.g. in one country or region.

6.3 Token based authentication

DC24 was the first opportunity to demonstrate JSON Web Token transfers at scale in the production system. Preparation had been two-fold: Middleware providers such as Rucio and FTS prepared software versions with a minimal set of features in time for DC24; in

		JINR		FNAL		IN2P3		RAL		PIC		KIT		CNAF	
Day	Scenario	DEST	SRC	DEST	SRC	DEST	SRC	DEST	SRC	DEST	SRC	DEST	SRC	DEST	SRC
1	T0 Export	1.42	N/A	1.13	N/A	1.09	N/A	0.76	N/A	1.18	N/A	1.16	N/A	1.17	N/A
2	T0 Export	1.46	N/A	1.12	N/A	1.10	N/A	0.50	N/A	1.17	N/A	0.94	N/A	1.17	N/A
3	T0Export, T1Export	1.31	0.62	1.08	0.88	1.33	1.03	0.72	0.99	1.18	1.06	1.10	1.06	1.28	0.93
4	T1 Export	N/A	0.37	N/A	0.91	N/A	1.12	N/A	0.76	N/A	1.05	N/A	0.95	N/A	1.00
5	T1-Export, Prod-out	1.18	1.72	1.15	0.87	1.25	0.89	0.98	1.01	1.21	1.09	1.23	0.77	1.17	0.77
6	T1-Export, Prod-out	1.14	2.42	1.18	0.88	1.47	0.88	0.72	0.81	1.17	1.03	1.19	0.76	1.18	0.95
7	T1-Export, Prod-out	1.19	2.19	1.15	0.87	1.22	0.87	0.81	1.04	1.20	0.98	1.21	0.73	1.16	1.02
8	AAA	1.30	N/A	N/A	1.10	1.39	N/A	1.31	N/A	1.31	N/A	1.70	N/A	1.32	N/A
9	All	0.38	0.34	0.87	0.84	0.57	0.57	0.95	1.02	1.25	0.86	0.86	0.56	0.65	0.25
10	All	0.70	0.34	0.98	0.74	0.58	0.65	0.56	0.99	0.70	0.66	1.03	0.98	0.63	0.28
11	All	0.63	0.33	0.91	0.73	0.43	0.76	0.77	1.05	1.09	0.84	0.91	1.09	0.69	0.24
12	All	0.40	0.54	0.92	0.86	0.89	1.00	0.85	1.15	1.21	0.87	1.13	0.89	0.78	0.29

Table 2. Ratio of the achieved throughput and the target rate for the different Tier-1 sites for each day of the challenge.

parallel CMS worked with the sites and storage middleware developers to get storage systems ready to cope with token authentication and authorization. The progress of the readiness was constantly monitored via SAM tests. CERN, all Tier-1 sites and 26 Tier-2 sites were ready to use tokens during DC24. Approximately half of all DC24 transfers were authorized via tokens, as shown in Figure 3.

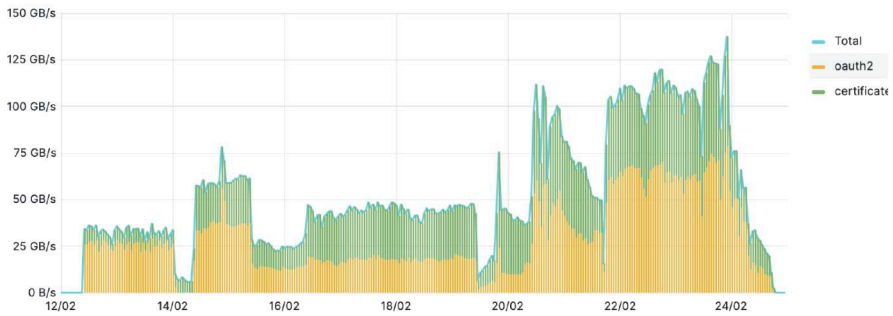

Figure 3. Transfer throughput during DC24 split by authentication and authorization method (X509 certificate or the new token).

7 Summary

The CMS experiment participated with the other LHC experiments ALICE, ATLAS and LHCb as well as the Belle II and DUNE experiments in Data Challenge 2024. Before the challenge four main network flows were identified and corresponding target rates for roughly 200 network links were determined. During pre-challenges the operations team became familiar with tools and monitoring, and the most important links were exercised. Therefore the 'minimal target' of 500 Gbs total throughput was reached without significant effort. In order to sustain the 'flexible target' of 1 Tbs various systems were found close to their limits and only careful tuning allowed the transfer rate to surpass the target successfully. During the challenge valuable experience was gained in operating Rucio and FTS under unprecedented load while tokens were used for authentication for the first time at production scale. Network bandwidth was found to be sufficient and limitations could be attributed primarily to components of the central data management tools or storage systems at the sites.

8 Acknowledgments

The CMS DC24 team thank the WLCG DOMA community for the fruitful collaboration during the entire DC24 including the preparation, the execution and aftermath. The intense exchange particularly with the ATLAS team led to a smooth operation of the dc_inject tool. The efforts from the middleware teams, most notably the FTS developers, were heroic and allowed the system to survive even under critical load. Site admins and network providers paid close attention to the infrastructure and reacted in timely manner.

References

[1] S. Chatrchyan et al. (CMS), The CMS Experiment at the CERN LHC, JINST **3**, S08004 (2008). 10.1088/1748-0221/3/08/S08004

[2] A. Hayrapetyan et al. (CMS), Development of the CMS detector for the CERN LHC Run 3, JINST **19**, P05064 (2024), 2309.05466. 10.1088/1748-0221/19/05/P05064

[3] I. Bird, LHC computing Grid. Technical design report (2005).

[4] dc_inject tool, https://gitlab.cern.ch/wlcg-doma/dc_inject

[5] M. Barisits, T. Beermann, F. Berghaus, B. Bockelman, J. Bogado, D. Cameron, D. Christidis, D. Ciangottini, G. Dimitrov, M. Elsing et al., Rucio: Scientific Data Management, Computing and Software for Big Science **3**, 11 (2019). 10.1007/s41781-019-0026-3

[6] K. Bloom, T. Boccali, B. Bockelman, D. Bradley, S. Dasu, J. Dost, F. Fanzago, I. Sfiligoi, A. Tadel, M. Tadel et al., Any Data, Any Time, Anywhere: Global Data Access for Science (2015).

[7] C. Collaboration, Tech. rep., CERN, Geneva (2021), this is the final version of the document, approved by the LHCC, https://cds.cern.ch/record/2759072

[8] S. Campana, WLCG data challenges for HL-LHC - 2021 planning (2021), https://doi.org/10.5281/zenodo.5532452

[9] D. Weitzel, D. Davilla, XRootD Monitoring Scale Validation (2021), https://doi.org/10.5281/zenodo.4688624

[10] A.A. Ayllon, M. Salichos, M.K. Simon, O. Keeble, FTS3: New Data Movement Service For WLCG, J. Phys. Conf. Ser. **513**, 032081 (2014). 10.1088/1742-6596/513/3/032081

[11] M. Lassnig, C. Wissing, WLCG/DOMA Data Challenge 2024: Final Report (2024), https://doi.org/10.5281/zenodo.11444180