EPJ Web of Conferences **245**, 04028 (2020) https://doi.org/10.1051/epjconf/202024504028
CHEP 2019

CMS data access and usage studies at PIC Tier-1 and CIEMAT Tier-2

Antonio Delgado Peris[2], *José* Flix Molina[1,2,*], *José M.* Hernández[2], *Antonio* Pérez-Calero Yzquierdo[1,2], *Carlos* Pérez Dengra[1,2], *Elena* Planas[3], *Francisco Javier* Rodríguez Calonge[2], and *Anna* Sikora[4]

[1]Port d'Informació Científica (PIC), Barcelona, Spain
[2]Centro de Investigaciones Medioambientales y Tecnológicas (CIEMAT), Madrid, Spain
[3]Institut de Física d'Altes Energíes (IFAE), Barcelona, Spain
[4]Universitat Autònoma de Barcelona (UAB), Barcelona, Spain

Abstract. The current computing models from LHC experiments indicate that much larger resource increases would be required by the HL-LHC era (2026+) than those that technology evolution at a constant budget could bring. Since worldwide budget for computing is not expected to increase, many research activities have emerged to improve the performance of the LHC processing software applications, as well as to propose more efficient resource deployment scenarios and data management techniques, which might reduce this expected increase of resources. The massively increasing amounts of data to be processed leads to enormous challenges for HEP storage systems, networks and the data distribution to end-users. These challenges are particularly important in scenarios in which the LHC data would be distributed from small numbers of centers holding the experiment's data. Enabling data locality relative to computing tasks via local caches on sites seems a very promising approach to hide transfer latencies while reducing the deployed storage space and number of replicas overall. However, this highly depends on the workflow I/O characteristics and available network across sites. A crucial assessment of how the experiments are accessing and using the storage services deployed in WLCG sites, to evaluate and simulate the benefits for several of the new emerging proposals within WLCG/HSF. Studies on access and usage of storage, data access and popularity studies for the CMS workflows executed in the Spanish Tier-1 (PIC) and Tier-2 (CIEMAT) sites supporting CMS activities are reviewed in this report, based on local and experiment monitoring data, spanning more than one year. This is of relevance for simulation of data caches for end-user analysis data, as well as identifying potential areas for storage savings.

1 Introduction

The estimates for the data volumes and computing required for the High-Luminosity LHC program (HL-LHC) show a major step up from the current needs. An intensive program of work was established by the WLCG[1] community to address this challenge. How to store the data collected and to ensure efficient access have been identified as the main challenges

*e-mail: jose.flix.molina@cern.ch

for HL-LHC, due to the increasing use of data volumes. Another challenge, from the site perspective, is due to the costs of operating and maintaining complex storage systems, with a mixture of disk and tape media storage services.

Half of the WLCG data is stored in Tier-1 centers, where massive data reduction campaigns are executed, producing reduced datasets for users. Most of the LHC experiments have enabled the possibility of accessing remotely files for specific processing activities, therefore some centers in WLCG have focused only on CPU operation, a trend which seems to grow in time. Reinforcing this trend by extending the possibilities for reliable remote access opens the possibility of a scenario where data can be concentrated in a reduced number of sites,and heterogeneous Computing facilities like HPC or Commercial clouds can be easily integrated. This strategy also presents the opportunity of orchestrating storage services as a single entity with common policies and quality of services (QoS) via a high level data management system, enabling new possibilities for storage resources optimization. These challenges are being addressed within the WLCG/HFS DOMA ACCESS[2] working group.

2 Methodology

Detailed studies about current usage and performance of storage resources in WLCG are needed if one wants to properly identify and evaluate or simulate the benefits for several of the new emerging proposals that are arising for storage re-organization.

The work presented here focuses on data access and popularity studies (number of accesses) from the storage accounting information at the Spanish Tier-1 (PIC) and Tier-2 (CIEMAT) sites supporting CMS activities. Both sites run dCache storage manager, and accounting records can easily been obtained from the dCache BillingDB database. The period selected for this analysis spans a year from September 2017 to September 2018 (the end of LHC Run2). These sites are located at ~10 ms RTT latency (Barcelona-Madrid), a sufficient low value for a consolidated/centralized storage deployment. In [3] we discuss dedicated studies done to understand how latency affects CMS workflows in the Spanish region.

3 Data access and usage studies

Typically, the disk-only storage at the Tier-1 and Tier-2 sites is constantly full. CMS has a service, the Dynamic Data Management (Dynamo) [4], which automatically deletes files from disk-only areas when placing new datasets that are going to be processed or allowing space for new processing campaigns (deletes or triggering transfers to Tape systems to allow for such space).

In the period considered for this study, the average disk utilization at PIC was ~2.3 PB, almost saturating all of the deployed disk for CMS. During the period, the disks were subject to a volume of ~9 PB of writes (10.5M files), ~24 PB of reads (3.5M distinct files), and ~9 PB of removes (11.0M files). This means that approximately the disk content was renewed every quarter, with an average rate of writes and deletions of ~25 TB/day. CMS Dynamo controls how the transfers and deletions are handled, according to some policies (sometimes you want to keep files on disk for a good reason), so the disk recycling is being dome in the most efficient way.

Typically, one would like to keep the most popular files on disk, so they are not deleted and re-placed in an unnecessary manner. One would like a fast access to occur as soon as new file is created on disk, and minimal time intervals since the last-access to the file and deletion time. Files which are rarely used in disk should be kept on tape systems, and re-called when needed.

3.1 Data Popularity

Data popularity studies for PIC and CIEMAT sites have been performed for the file accesses occurred in the selected period, by data types, for both experimental data (categorised as DATA in the following sections) and *simulation* (referred to as Monte Carlo, or MC) CMS files. A careful analysis of file accesses is essential to identify those files that could eventually be cached in the Spanish region. Many data files are accessed remotely from the compute nodes at PIC and CIEMAT. Files read from, outside the PIC-CIEMAT federation, are not recorded into either of the storages databases, therefore are not included in this study. However, given the global scope of CMS operations, it is safe to assume that data access patterns are similar for data reads and exports, therefore popular files, as identified from our storages, can be assumed to be popular elsewhere across the CMS grid.

Figure 1 presents an example of a popularity plot for MC AOD files at CIEMAT. The plot shows that AOD files are really popular in the site, since the average access is ∼15 (this is the average number of accesses from creation to deletion). CMS Tier-2 sites execute mostly analysis jobs, whose input files are typically AODs (in our study, AOD type include AOD, MiniAOD and NanoAOD formats). In particular, many users run over the same datasets, and many analysis are repeated and refined. This explains why these files are popular. Even if PIC Tier-1 also exports MC AOD files to tasks running at T2 sites, RAW MC files are read much often, since AOD datasets derive from the same RAW MC files, and also they include pile-up samples. The table 1 summarizes the popularity results for the diverse data types at PIC and CIEMAT. Differences between a Tier-1 and a Tier-2 are can be clearly appreciated in terms of storage composition and relative popularities of the respective samples.

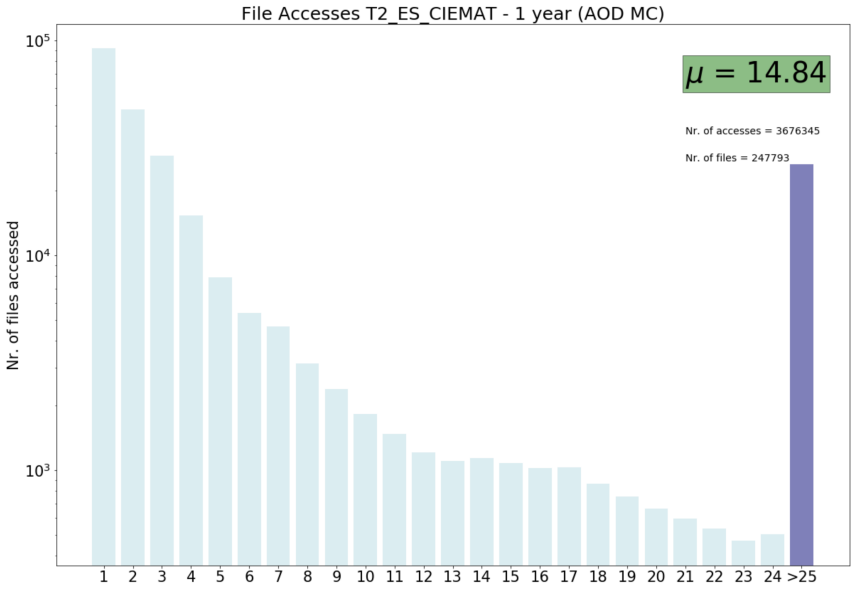

Figure 1. Data popularity distribution (in number of times a given file is accessed, for all the files present in the site catalogue during the period of study) for MC AOD files at CIEMAT Tier-2 site (1-year period).

Table 1. Data popularity measurements by data type both at PIC and CIEMAT sites (1-year period).

	PIC Tier-1			CIEMAT Tier-2		
DATA files	RAW	RECO	AOD	RAW	RECO	AOD
Nr of files	248,686	102,947	432,134	56,401	78,221	327,737
Nr of accesses	968,810	201,803	1,461,973	158,542	284,552	3,235,255
Avg. accesses	3.90	1.96	3.38	2.81	3.64	9.87
(95% percentile) accesses	~16	~3	~10	~8	~10	>25
	PIC Tier-1			**CIEMAT Tier-2**		
MC files	RAW	RECO	AOD	RAW	RECO	AOD
Nr of files	76,776	60,967	366,173	55,710	32,116	247,793
Nr of accesses	1,464,607	253,235	1,712,961	199,910	705,493	3,676,345
Avg. accesses	19.08	4.15	4.68	3.59	21.97	14.84
(95% percentile) accesses	>25	~4	~11	~10	>25	>25

3.2 Data lifetime

Files are temporarily placed in disk areas at PIC and CIEMAT. For the selected period at CIEMAT, approximately half of the DATA files of all types (RAW, RECO or AOD) were deleted within a month. At PIC, half of AOD DATA files are deleted within 10 days, and RAW DATA files are typically left on disk longer. Only 30% of RAW DATA files are deleted within a month. These files are usually kept on disk on purpose, since processing campaigns might need to re-process data, and it is more convenient to keep files on disk rather than restoring them from tape systems.

Figure 2 shows the lifetime of MC files at PIC and CIEMAT, demonstrating that MC files live longer at the Tier-2 than the Tier-1, where the space is constantly renewed to accommodate new processing campaigns. Again, in both cases, half of MC AOD files are deleted within 10 days. The AODs are produced in PIC, cleared when the processing is finished, and in the CIEMAT they are analyzed, and they remain there while they are popular. Average lifetime values are provided, and it's worth mentioning that average lifetimes are about 50% longer at CIEMAT than PIC for all MC datasets.

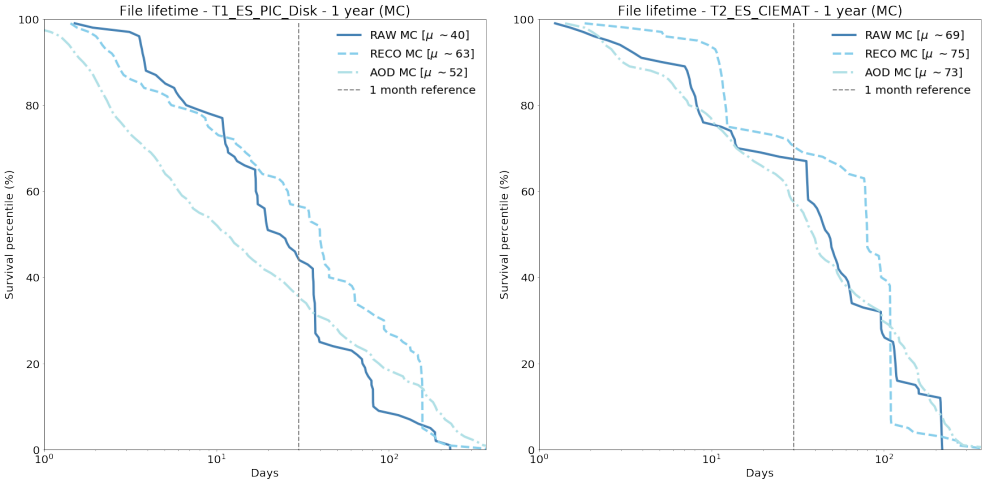

Figure 2. Cumulative distribution, or lifetime ($t_{deletion} - t_{creation}$) of MC files by files type at PIC [left] and CIEMAT [right] (1-year period).

3.3 Data access

During the adventurous lifetime of a file, several accesses might occur. For the selected period and file types, we studied the typical time intervals between file creation and first access, at both sites. Looking at DATA files, approximately 20% are firstly accessed in their first day since creation, in both the Tier-1 and Tier-2, while around half of them are first accessed within 10 days. The pattern is similar at both sites.

Figure 3 shows the time between file creation and first access for MC datasets at PIC and CIEMAT. Half of RAW MC files are accessed within a day in PIC, since they are popular samples or they have been transferred or produced at the site for a campaign. The same occurs at CIEMAT with AOD MC and RECO AOD files. As expected, popular files have a prompt access, compared to files that are not so popular. It can be seen that RAW MC files stored at CIEMAT are not very popular, since half of them are created and not accessed at all within the first month. On the contrary, for the other data types at both sites, approximately 80%-90% of files have been all firstly accessed within a month.

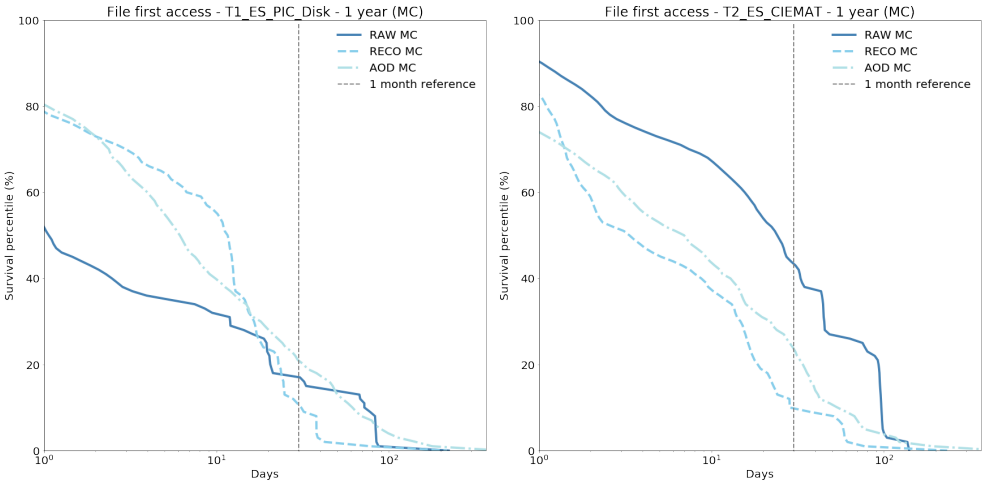

Figure 3. Time lapse between creation and first access for each type of MC dataset at PIC [left] and CIEMAT [right] (1-year period).

Popular files are read often, and the which type of datasets are popular might be different at Tier-1 or Tier-2 sites. Figure 4 shows the time intervals between file accesses for DATA dataset types, for both PIC and CIEMAT, and in the period analyzed. Approximately 90% of RAW DATA re-reads happen within a day, since many jobs read the same input file in a processing campaign (job splitting). It can be observed that most of the files are re-read within a month, and there is a clear pattern for different data types. Popular files are read much often, hence the time between re-reads is small. Similar distinctive patterns are seen for the three MC dataset types at both sites.

The time interval between last access and deletion from storage is as well an important parameter to monitor. RAW DATA at PIC is kept on disk longer than other data types, since last access. Half of the RAW DATA has last access to deletion time >1 month. For the rest of data types (RECO and AOD), ~25% of files have last access to deletion >1 month. At CIEMAT we observe that popular MC AOD samples are deleted promptly from last access, since 50% of the files have a last access to deletion time interval shorter than 10 days.

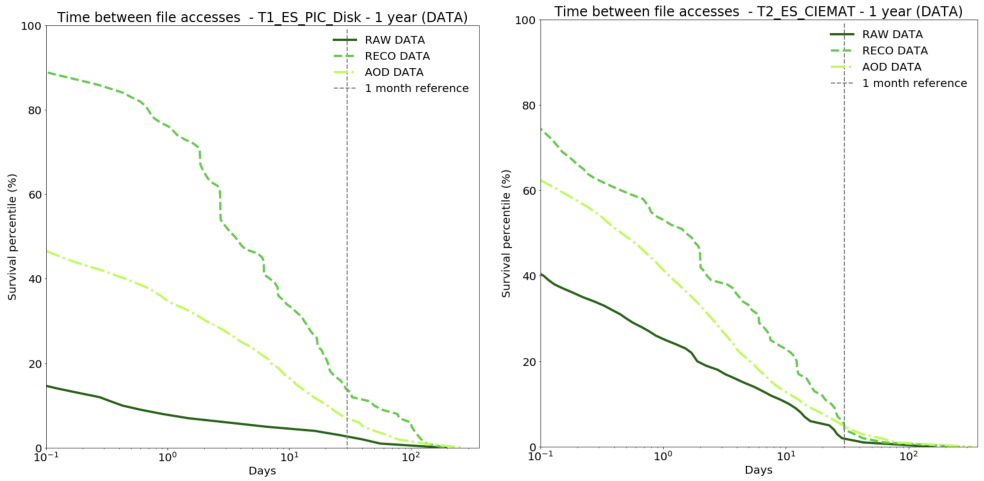

Figure 4. Time interval between successive file reads for DATA files at PIC [left] and CIEMAT [right] (1-year period).

3.4 Data redundancy

We constantly check how much data is redundant in the region (for Tier-1 and Tier-2 levels). Data redundancy in the region is very low, which means no significant gain in disk volume could be derived from a storage consolidation, with no further optimizations. For example, CMS dataset blocks comparisons for PIC and CIEMAT on 30th October 2019 showed that out of 122.58 Kblocks (1.96PB) at PIC, and 56.17 Kblocks (1.84PB) at CIEMAT, only 2.23 Kblocks were present at both PIC and CIEMAT, with a total size of ~10TB, which is <0.5% of the storage available at the sites. Typically, the data redundancy in the region is below 5%.

4 Outlook and conclusions

These research activities are aimed at providing input towards a more efficient deployment scenarios and techniques to alleviate the gap between the expected storage resource needs and the flat budget technical evolution towards HL-LHC. There is a rich resource of data concerning disk usage present at the storage accounting systems, along with data from all the computing tasks executed, for each of the sites. This information can be extremely useful to understand where to focus, and in particular to model how new scenarios would behave once deployed.

The preliminary studies presented in this paper are aimed at understanding how the storage systems are utilized at both PIC Tier-1 and CIEMAT Tier-2 by CMS. These studies have been centered on data access and popularity patterns. Next steps include:

- the study of data accesses by protocol, since GridFTP is used for data transfers (copy ot move) between storage endpoints (FTS), while data access by CMS jobs are done via XRootD.
- the detailed study of latency read effects in detail. Accounting from CMS jobs executed in the region will be studied, focusing on those with remote reads from well known and distant sites.

- remote XRootD reads showed file transfer volumes exceeding file sizes. This represents around 10% of the total exported volume. This showed up during the study and needs further understanding.
- the modelization of the access patterns and identify which data is susceptible to be cached and which could be the benefits in the region (performing simulations based on real data accesses from CMS payloads).
- an expansion of our analysis to include the usage of tape system and tape buffers at PIC.
- an extension of the study to include CERN, in order to cover all Tier levels in their use of CMS data.

This work was partially supported by MICINN in Spain under grants FPA2016-80994-C2-1-R, and FPA2016-80994-C2-1-R, which include FEDER funds from the European Union.

References

[1] WLCG project: http://wlcg.web.cern.ch/
[2] X. Espinal, et al, *The Quest to solve the HL-LHC data access puzzle. The first year of the DOMA ACCESS Working Group*, to be published in these proceedings.
[3] A. Delgado, et al, *Lightweight site federation for CMS support*, to be published in these proceedings.
[4] Dynamo - Dynamic Data Management System: https://ddm-dynamo.readthedocs.io/