PIC
port d'informació
científica

Simulating a content delivery network solution for the CMS experiment in the Spanish WLCG Tiers

Carlos Pérez Dengra on behalf of CMS Collaboration

ISGC 2022: International Symposium on Grids & Clouds 2022 Virtual Conference, 20-25 Mar 2022

IFAE 25 years 1991-2016 Institut de Física d'Altes Energies

EXCELENCIA SEVERO OCHOA

BIST Barcelona Institute of Science and Technology

GOBIERNO DE ESPAÑA MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD

Ciemat Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas

EXCELENCIA MARÍA DE MAEZTU

cfp CIEMAT física de partículas

PIC
port d'informació
científica

Contents

-Introduction.

-Research on data popularity of CMS.

-Simulating caches at PIC and CIEMAT.

-Summary results.

-Outlook.

PIC
port d'informació
científica

Introduction

Since 2009 LHC have stored more than 1 Exabyte of simulated and collision data in disk and tape. Proton and ion collisions will rise up to a factor 10 (as compared to today's values) at the HL-LHC era in 2029. LHC scientists work on an ambitious R&D program to fit the future compute resources requirements onto the available compute funding budget.

Data Lake model is one of the proposed changes in the infrastructure.

→ Consolidate storage resources in fewer big sites.

→ Deploy cache systems to bring data close to compute resources.

Spanish CMS sites are currently exploring these new mechanisms: content delivery network (CDN) solutions and their effects on executed tasks efficiencies.

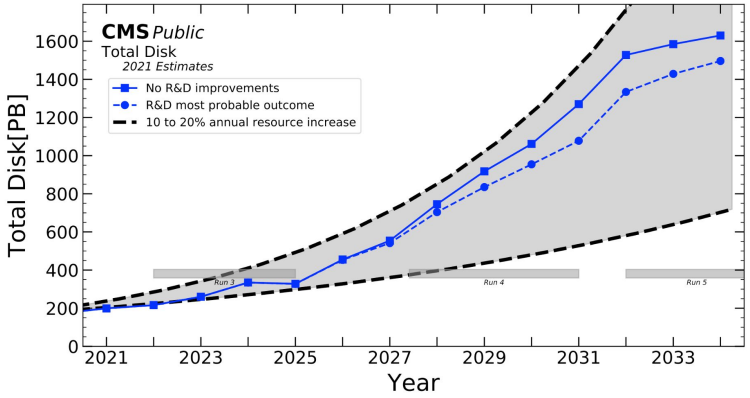


Figure 1.- Extrapolated 2021 estimates for expected collision and Monte Carlo data to be stored at disk drive CMS resources[1].

[1] CMS Offline and Computing Public Results: Approved HL-LHC resource projections (November 2021): https://twiki.cern.ch/twiki/bin/view/CMSPublic/CMSOfflineComputingResults, Consulted on 14th February of 2022.

PIC
port d'informació
científica

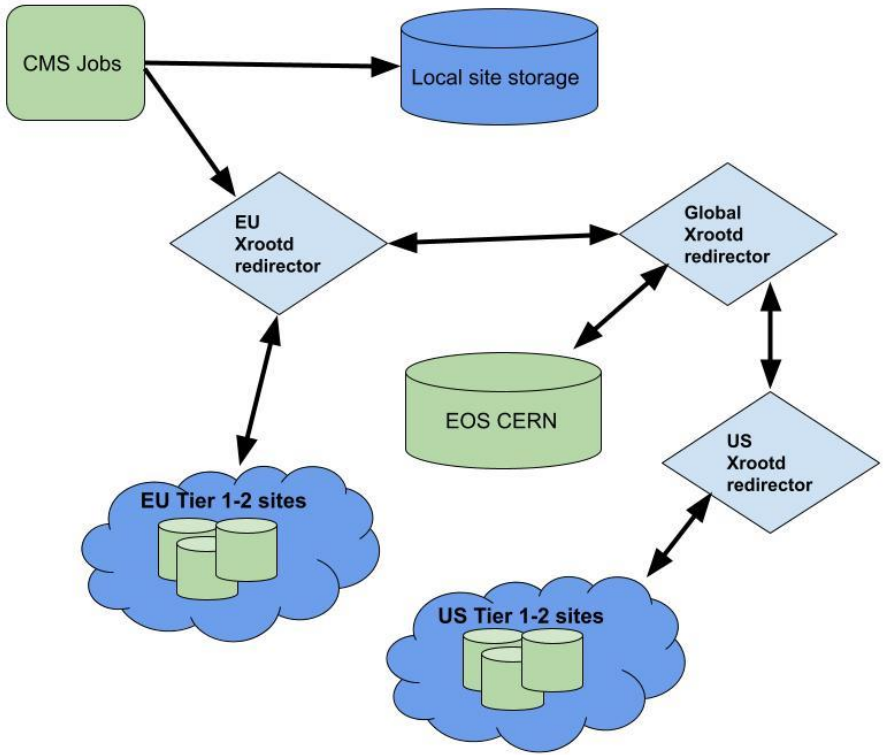
Introduction

Why do we want to deploy caching systems?

CMS tasks are generally executed at the computing site storing their input datasets in order to ensure local data access.

However, using a Xrootd redirectors engine, jobs running at a particular site can also access remotely any data required, no matter where in the world they are. This service is called AAA (Any Data, Anytime, Anywhere)

Deploying caches would bring data close to compute nodes when required by jobs, instead of being remotely accessed.



PIC
port d'informació
científica

Introduction

Spanish CMS sites of PIC Tier-1 and CIEMAT Tier-2 are currently exploring CDN solutions based on the deployment of caches.

The inclusion of caches in the data access architecture from the Spanish CMS region aims to:

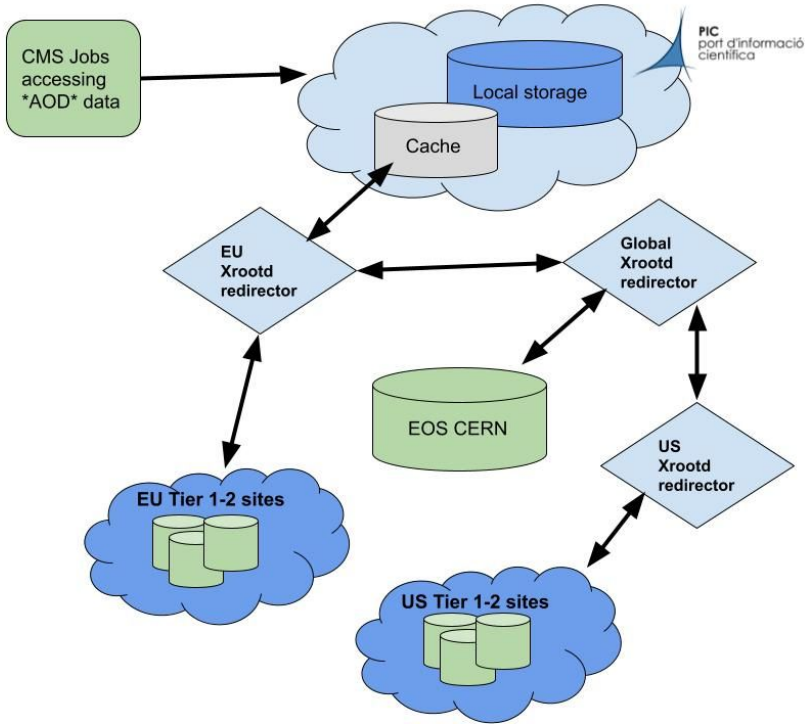
- **Reduce data access latencies.**
- **Improve the global CPU efficiency.**
- **Reduce bandwidth.**
- **Reduce storage costs.**

In this contribution the expected behavior of LRU caches deployed in the spanish CMS region using real data is simulated. CMS Software (CMSSW) jobs report to Popularity database, where the information of data accessed is stored. Simulations will allow the sites to predict and measure the impact of deploying this caching system in our sites and explore the optimal configurations.

Research in data popularity of CMS data

In the different CMS workflows, the jobs access different types of data: the **CMS Data tiers**. The Data tiers are the categories that receive the collision and experimental data according to their degree of processing and content (e.g. detector hits vs particle tracks and vertices).

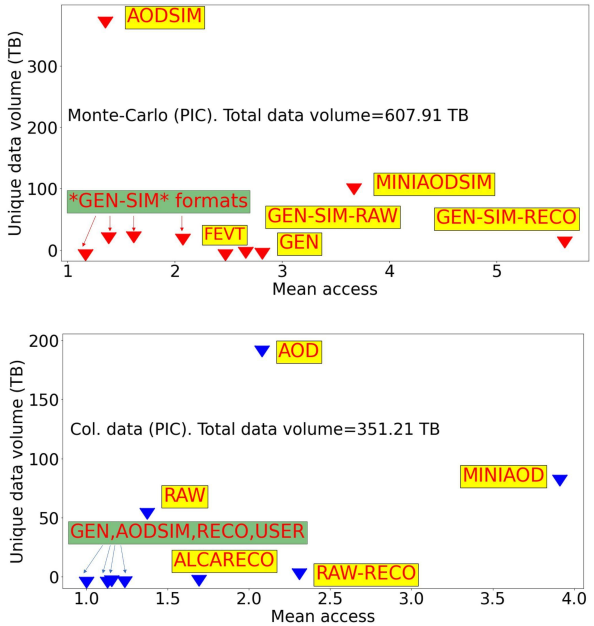
Good performance of the cache involves storing data that is highly accessed (**popular**). One can evaluate which data tiers are most accessed by the jobs in order to filter them and increase their CPU efficiency.

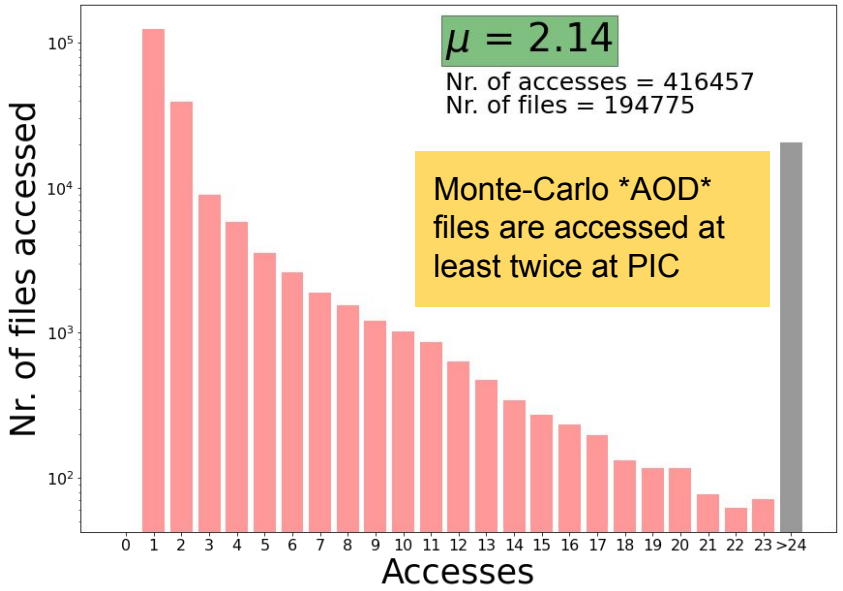


Research in data popularity of CMS data

The selection of data tiers being cached is based on their popularity (mean access of file by data tier) and their total unique volume data (the total volume accessed of unique files).

Results showed that all AOD data (*AOD*) is the preferred data tier to be cached.





PIC
port d'informació
científica

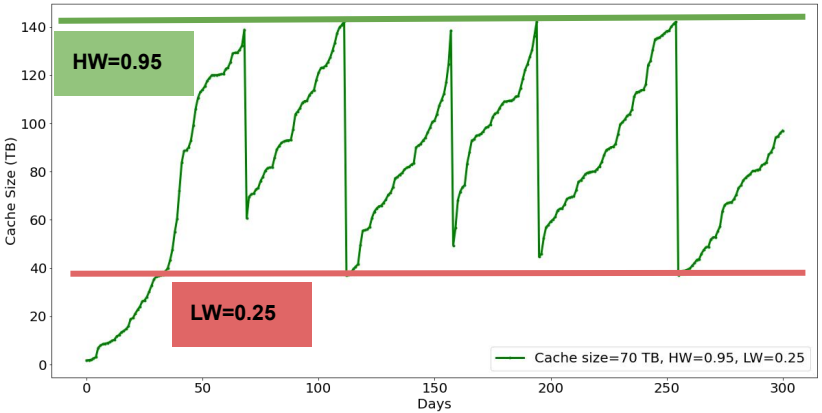
Simulating caches at PIC and CIEMAT

Cache modellization:

1.- Cache fills up by adding unique files accessed by jobs in chronological order (as a physical cache would do).

2.- When the cache is full, we apply a Least Recently Used (LRU) algorithm: organize files by order of use and inclusion date. The ones not accessed in the largest period of time are deleted.

3.- Deletion process based on watermarks: thresholds that indicate a certain range of cache occupancy. If the cache occupancy overcomes "a high watermark HW", the last accessed files are removed until reaching "a low watermark LW".



PIC
port d'informació
científica

Simulating caches at PIC and CIEMAT

During the simulation, the following metrics are computed to evaluate the expected cache performance:

$$HR = \frac{hits}{hits + misses} = \frac{hits}{N_{acc}}$$ **Hit-rate**: where ***hits*** are the number of times a file required by jobs during the whole analyzed period is already in cache and ***misses*** not (and has to be cached).

$$\frac{Mean\ hits}{month \cdot files}$$ **Monthly mean number of hits by file**: the survival data in the cache is the most popular. Consequently, if data in cache is very popular during the period, the average number of hits per month and per file increases over the time and stabilizes. This value should improve at certain point the mean popularity value itself.

PIC
port d'informació
científica

Simulating caches at PIC and CIEMAT

Conditions and constraints:

1.- Files accessed by experiment tests are excluded from the analysis.

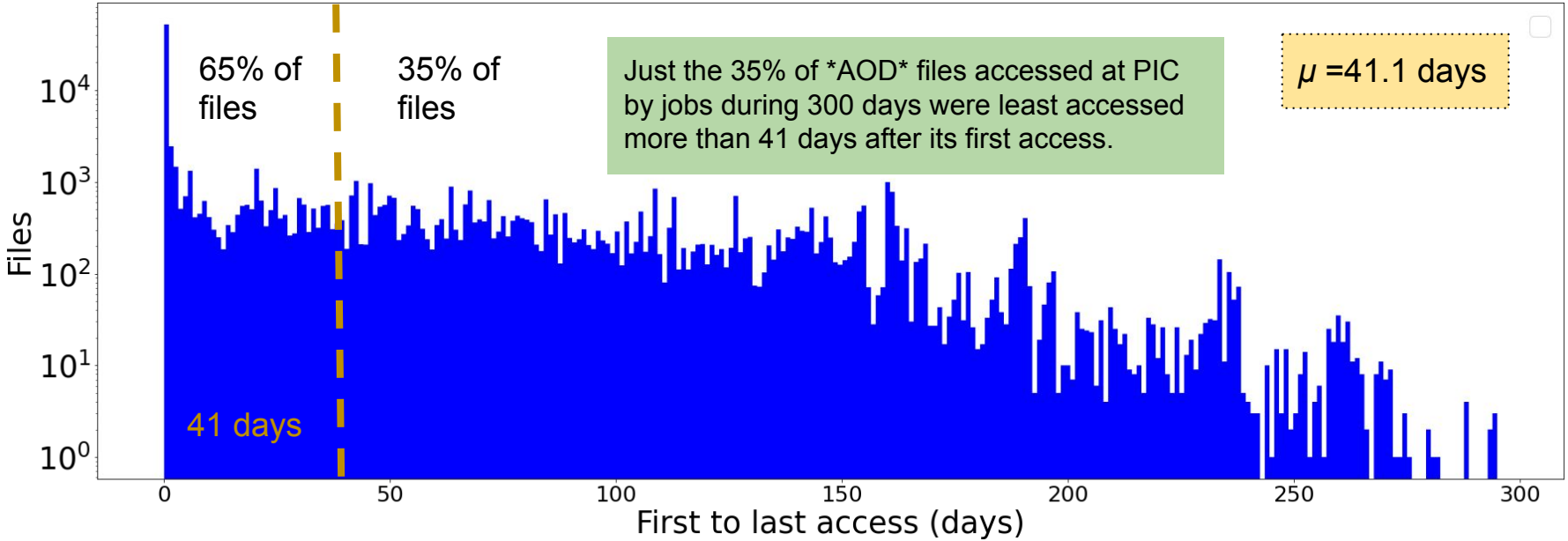
2.- Intermediate experimental data kept at the storage for short periods of time (days), called unmerged data, and intermediate files placed in execution nodes are also excluded.

3.- This simulation does not take into account whether the files are already at the site or not. This approach assumes that data required by jobs would be all cached instead of being written at local storage.

4.- The input caches sizes have been estimated to fill the cache between 50-60 days (about 2 months).

Simulating caches at PIC and CIEMAT

Last to first access to *AOD* files Collision and Monte-Carlo data @ PIC

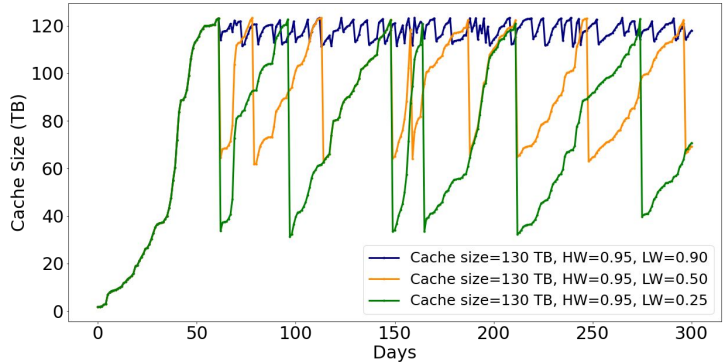


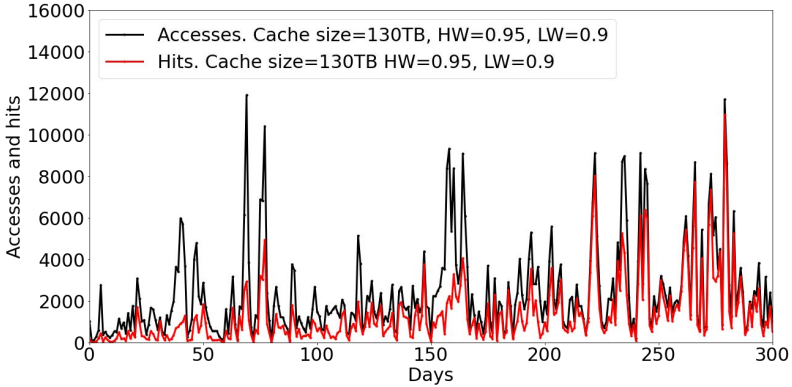
The result agrees with the one obtained by the studies carried out on the PIC local storage with dCache (42 days) **[2]**. This shows that data accessed by the jobs have a usage time similar to that in the storage system of the sites.

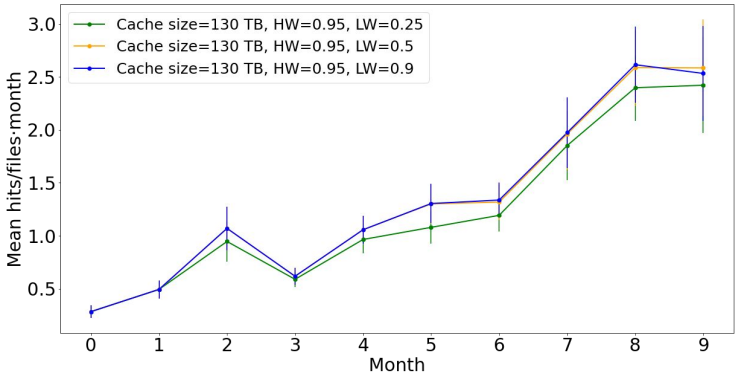
[2] Delgado Peris, A., Flix Molina, J., Hernández J., Pérez-Calero Yzquierdo, A., Pérez Dengra, C., Planas, E., Rodríguez Calonge, J., Sikora, A 2019 "CMS data access and usage studies at PIC Tier-1 and CIEMAT Tier-2", EPJ Web Conf., 245 (2020) 04028

Simulating caches at PIC and CIEMAT

PIC CMS Tier-1 @ XCache for *AOD* files only (130 TB of cache size).





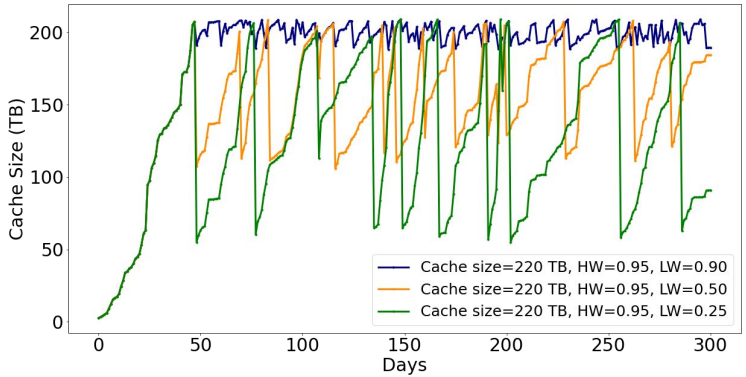


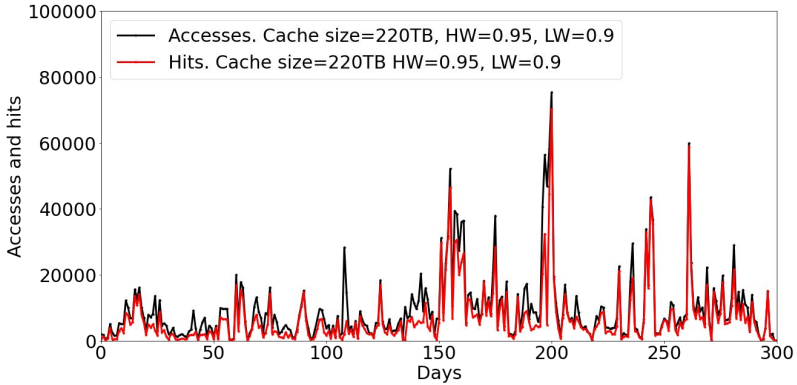
-The best results have been shown to be with LW=0.9.

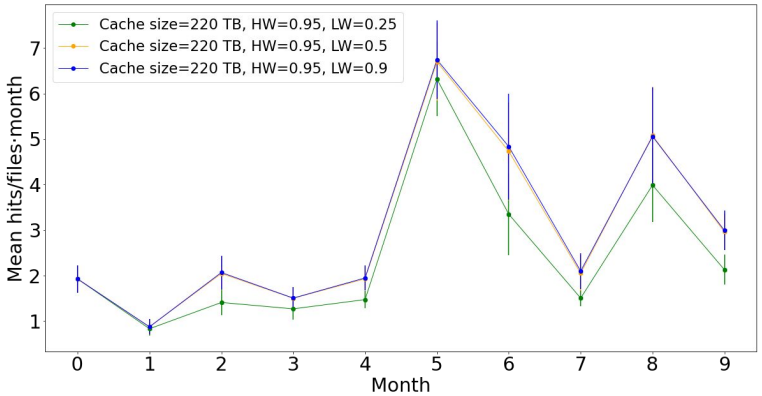
-The average hit-rate value during the last 30 days was 0.77, with a maximum value of 0.9, with a total of 3.3E5 accesses during the period.

Simulating caches at PIC and CIEMAT

CIEMAT CMS Tier-2 @ XCache for *AOD* files only (220 TB of cache size).



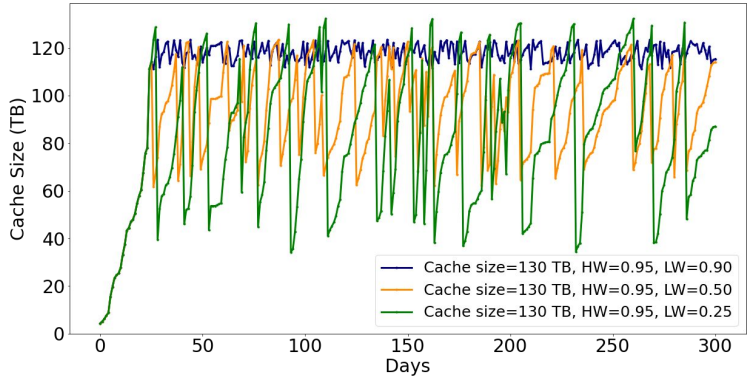


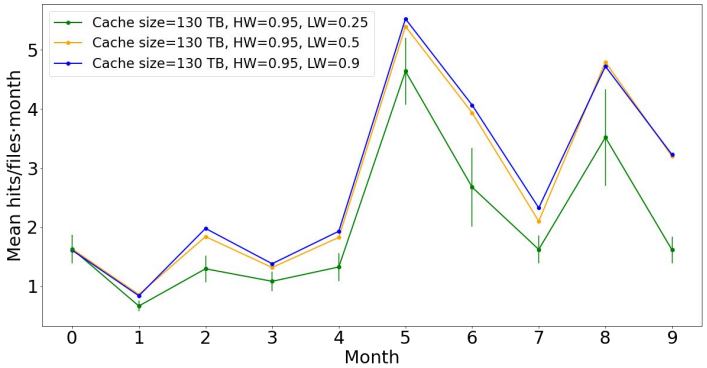


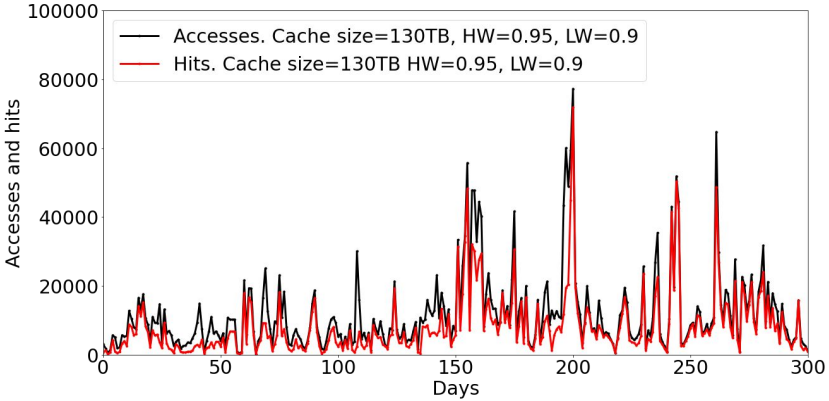
- CIEMAT has a higher rate of access to data (9.2E5) and higher mean of last to first access to files (67.9 days).
- The average hit-rate value during the last 30 days have been 0.8, with a maximum of 0.92.

Simulating caches at PIC and CIEMAT

PIC+CIEMAT CMS Tier-2 @ XCache for *AOD* files only (130 TB).





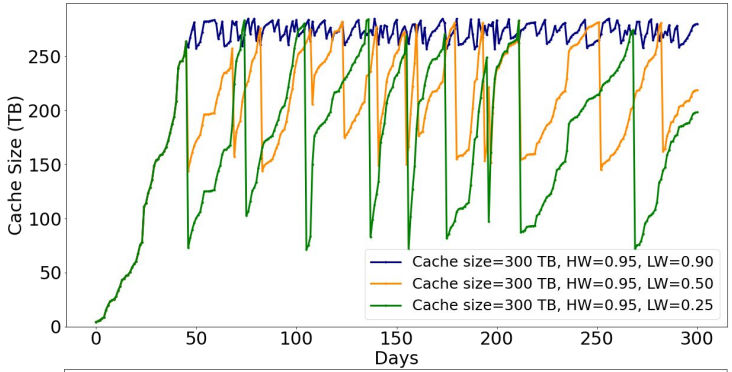


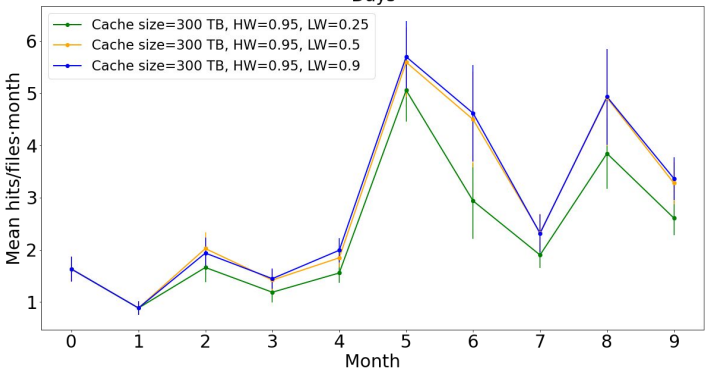
-This simulation is based in the scenario where the cache resources would be shared by PIC and CIEMAT and located at PIC Tier-1 (100 Gbps bandwidth - 10 ms of latency between both sites).

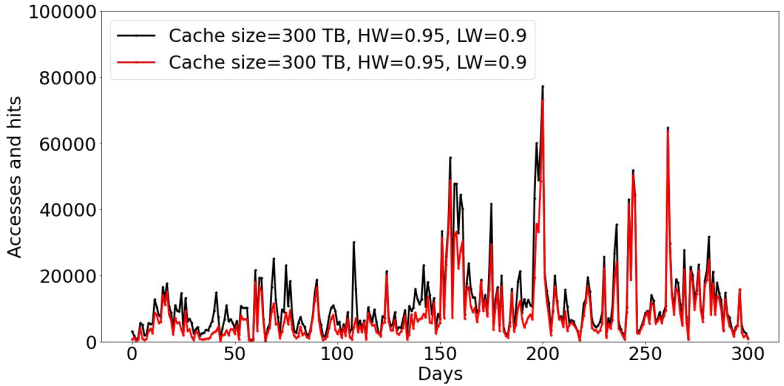
-The hit-rate during the last 30 days has reached a value of 0.82, with a maximum of 0.93 during the period.

Simulating caches at PIC and CIEMAT

PIC+CIEMAT CMS Tier-2 @ XCache for *AOD* files only (300 TB).







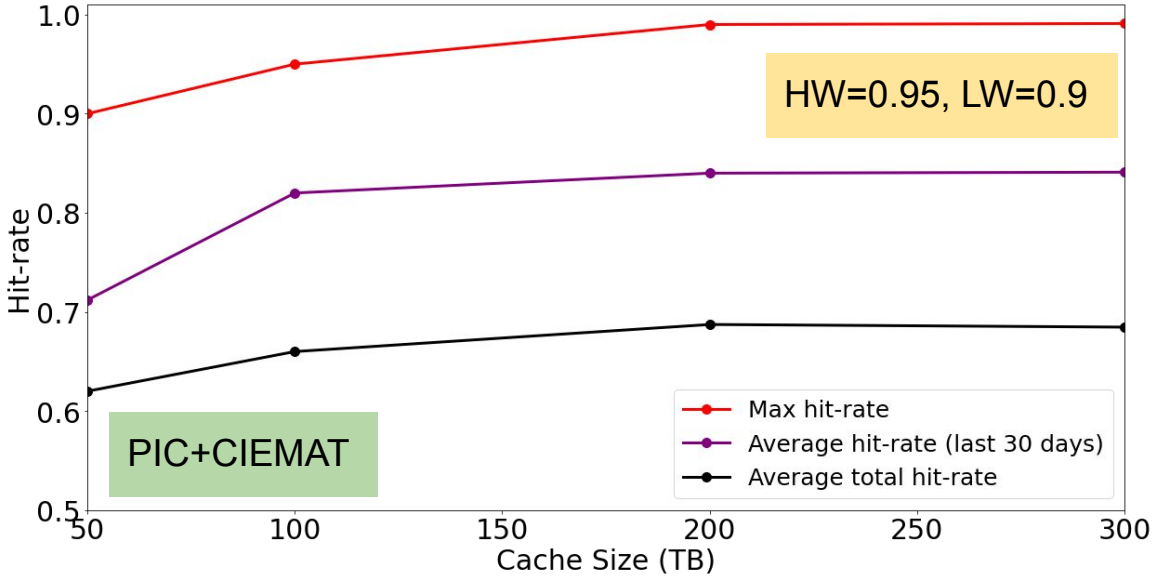
-Expanding even more the available cache size (up to 300 TB) we reach an average hit-rate of 0.84 in the last 30 days and maximum hit-rate of 0.99, being the first configuration to achieve this value.

Summary results

-The relevant hit-rate metrics in this simulation are computed for PIC and CIEMAT merged accesses to *AOD* files during 300 days.

-Increasing the cache size from 50TB to 200TB has a 12% improvement in the average hit-rate over the last 30-day period. From that value of the cache size, both in this and the rest of the metrics, the value stabilizes. This is because the use of a restricted data window of 300 days in the simulation does not allow the value to evolve further. This value has to ideally go beyond this value by performing a simulation over 300 days.

-The maximum daily hit-rates of 0.99 also level off at 200TB cache size.



PIC
port d'informació
científica

Summary results

	PIC	CIEMAT	PIC+CIEMAT
Accesses	3.3e5	9.2e5	1.2e6

PIC CMS Tier-1

HW=0.95, LW=0.9

Cache size (TB)	Mean HR (Last 30 days)	Max HR
130	0.77	0.9

CIEMAT CMS Tier-2

HW=0.95, LW=0.9

Cache size (TB)	Mean HR (Last 30 days)	Max HR
220	0.8	0.92

PIC +CIEMAT

HW=0.95, LW=0.9

Cache size (TB)	Mean HR (Last 30 days)	Max HR
130	0.82	0.93

Cache size (TB)	Mean HR (Last 30 days)	Max HR
300	0.84	0.99

The results of the simulation show that the proper cache size for PIC and CIEMAT oscillate between 200 and 300TB for a LRU cache simulation of 300 days.

PIC
port d'informació
científica

Outlook

-Simulations have been carried out in the PIC CMS Tier-1 and CIEMAT Tier-2 sites modeling the behavior of the cache using real data access to the data through CMSSW Popularity database. These simulations allow PIC and CIEMAT sites to understand the expected behavior of deploying caches in CMS sites in the spanish region. Also, this approach allow the sites to estimate the proper sizes that caches should have according to the rate of data access.

-Although the model may include improvements (taking into account whether the files are already located locally in the storage or not), these results show the possibility of using a shared caching system between PIC and CIEMAT between 200-300TB to cover the accesses to *AOD* data. This would be possible thanks to the low latency between the two sites (10 ms). These results could be also applicable to other CMS Sites.

-Upcoming tasks will include comparing the results with XCache nodes (the chosen cache service using xrootd) in test deployed in PIC Tier-1 and CIEMAT Tier-2 to evaluate the possible compute, storage and bandwidth gains. The results obtained in theses simulations will be applied in the future XCache service in production in both PIC and CIEMAT

Thank you for your attention

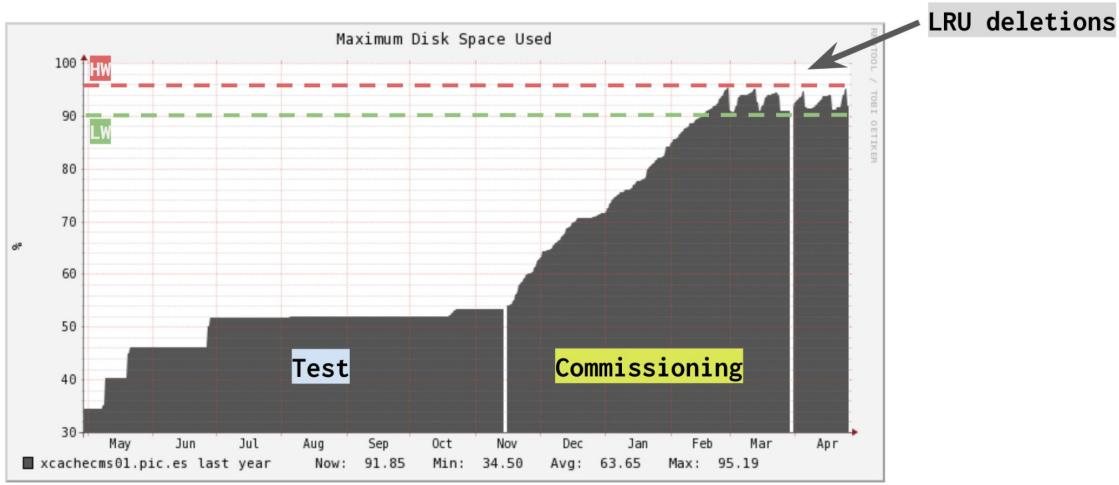
PIC
port d'informació
científica

Backups

Current caching status at PIC&CIEMAT

In May 2020, a 130 TB test XCache node was deployed in the Tier-1 of the PIC and 20 TB in the CIEMAT.

Both sites are configured to cache *AOD* files based on their popularity. All files that are not on the site ('fallback' mode), are served and downloaded to the node. Data retention policy in the cache is based on LRU (Least Recently Used), that deletes the least accessed data when the cache is full.



Year	Granted	Used & forecasted
2021	200 TB	150
2022	200 TB	200
2023	200 TB	200
2024	200 TB	200
2025	200 TB	200