PIC
port d'informació
científica

PIC: Storage studies for CMS

Carlos Pérez Dengra

HSF WLCG Virtual Workshop 2020, 20th Nov

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

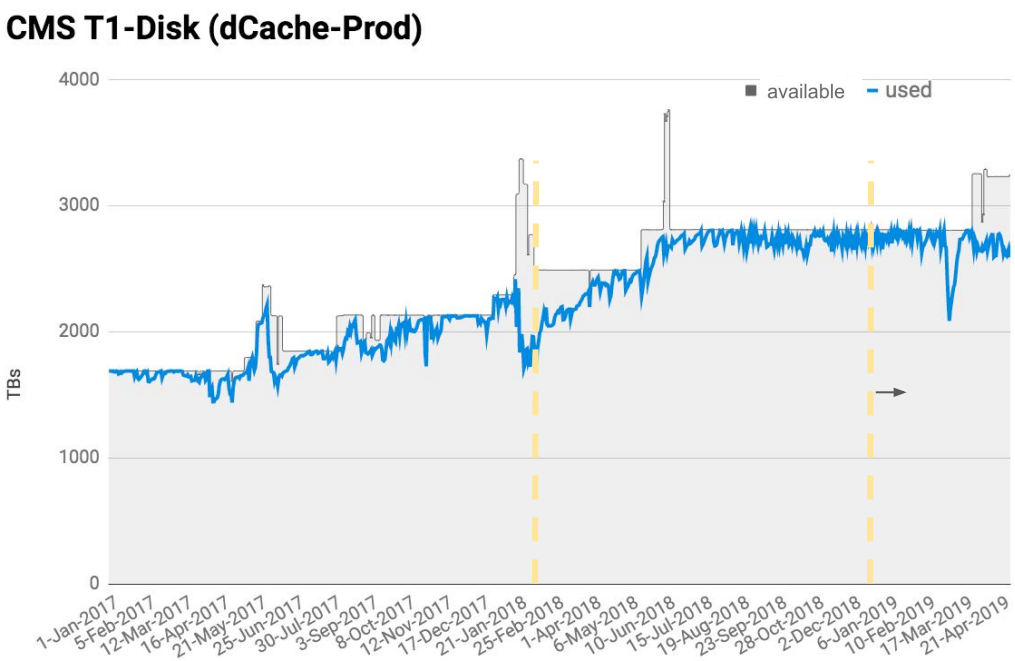
PIC
port d'informació
científica

PIC storage at a glance

- **PIC is a Tier-1 in WLCG (~5% of resources) and supports a variety of disciplines**
- **10 PB disk on dCache 5.2.30**
  → dCache pools in dual-stack
- **32 PB tape on Enstore 6.3.4-2 (CentOS7)**
  → Technology: T10KC/T10KD (STK8500 library) and LT08 (new IBM TS4500 library)
- **Active participation in DOMA activities:**
  → EULake: EOS server deployed at PIC (~60 TB) and integrated into its testbed
  → TPC enabled for HTTPs and XRootD (included in TPC DOMA testbeds)
  → Token authentication enabled for wlcg VO
  → Latency effect studies for CMS workflows in the Spanish region and others
  → Storage access and popularity studies based on local dCache instance and information from the CMS jobs
  → xCache deployed (~150TB) to understand the service (to be expanded soon)
- **Spanish CMS sites separated at ~10 ms latency, suitable for a common storage future solutions and services**

PIC
port d'informació
científica

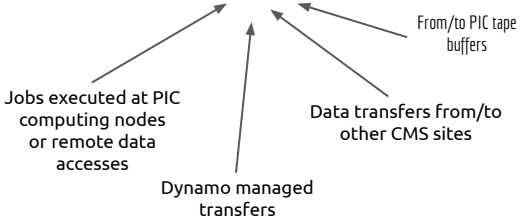
CMS disk utilization (2017-2019)



CMS saturates and utilize all of the available disk at PIC

As example of 2018...
Average disk utilization ~2.3 PB

9 PB writes (10.5M files)
9 PB removes (11.0M files)
24 PB reads (3.5M distinct files)



Latency effects on CMS Workflows

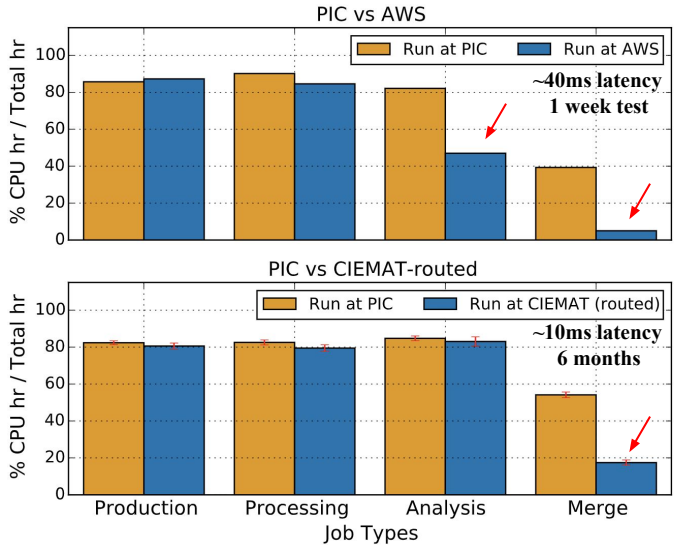
- **Effects on the efficiency of jobs in remote compute nodes**

  → Making use of the existing CMS xrootd federation infrastructure, HTCondor re-route of ~5% of jobs between PIC Tier-1 (Barcelona) and CIEMAT Tier-2 (Madrid)

  → Cloud bursting tests (AWS) - data center at Frankfurt (40 ms of latency). HTCondor-CE modified to send CMS jobs to Amazon nodes. Not so good for I/O intensive jobs

Filtering those jobs reading data from PIC

See CHEP'19 contribution



Jobs run at PIC - reading from PIC
Jobs routed to AWS - reading from PIC

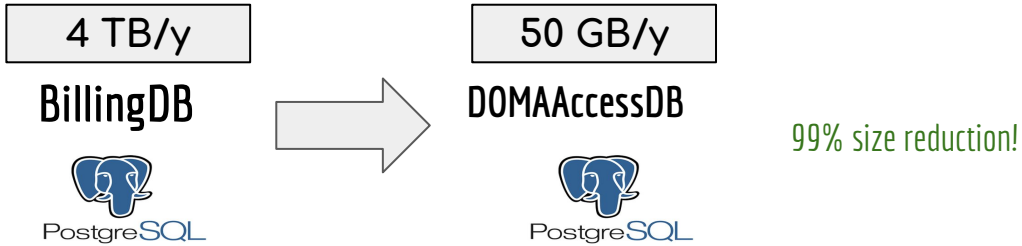
Jobs run at PIC - reading from PIC
Jobs routed to CIEMAT - reading from PIC

PIC
port d'informació
científica

Data accesses studies from dCache

→ All the of accounting information at PIC and CIEMAT is available at the dCache **billingDB**

→ A **new and smaller Postgres BBDD** has been generated to hold all the relevant information for the CMS data accesses for further analysis

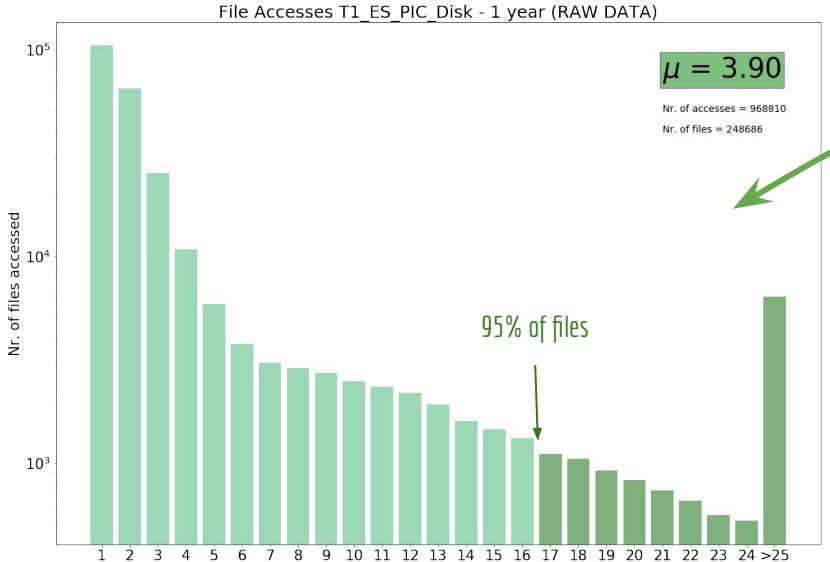


→ Data analysis and calculations are performed spawning Jupyter Notebooks to the PIC farm (HTCondor)

See CHEP'19 contribution

PIC
port d'informació
científica

Data popularity from dCache 1/2



	DATA					
	PIC			CIEMAT		
	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	3.9	1.96	3.38	2.81	3.64	9.87
(95% percentile) accesses	~16	~3	~10	~8	~10	>25

Popular Data Tier

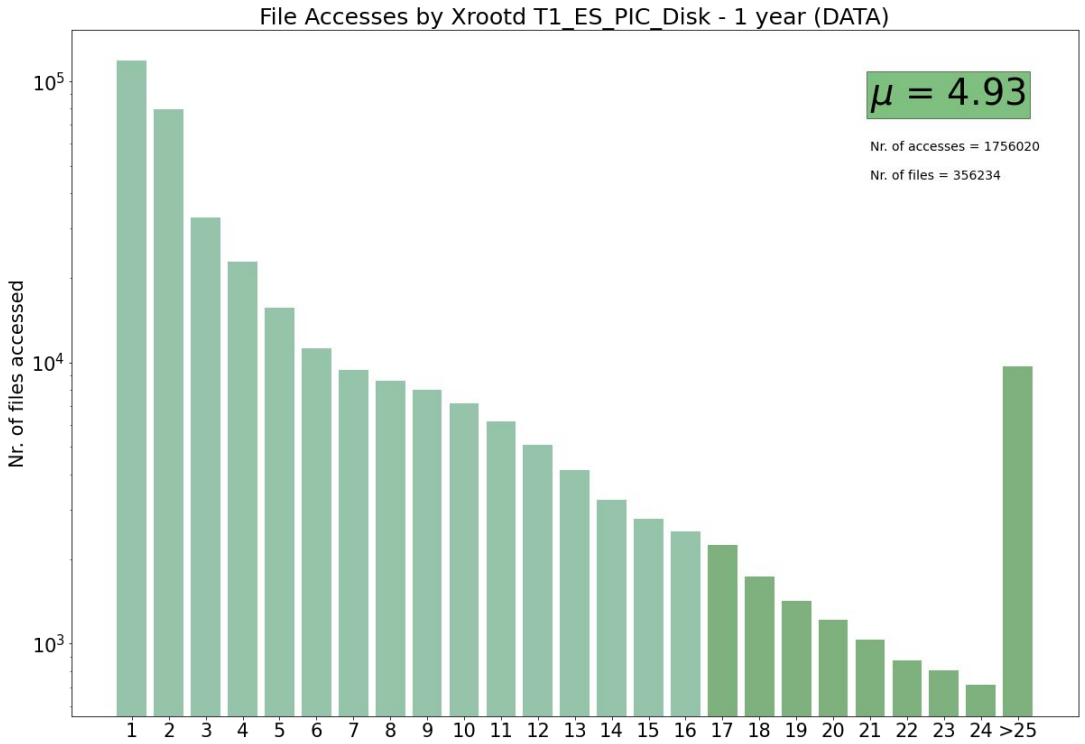
DATA RAW → RAW
DATA RECO → ALCARECO, RAW-RECO, RECO
DATA AOD → AOD, MINIAOD, NANOAOD

Popularity in storage allow us to understand which amount of data are re-accessed (or never) in our disk drive servers by data tier and protocol.

PIC: ~3.4 accesses/file
CIEMAT: ~8.0 accesses/file

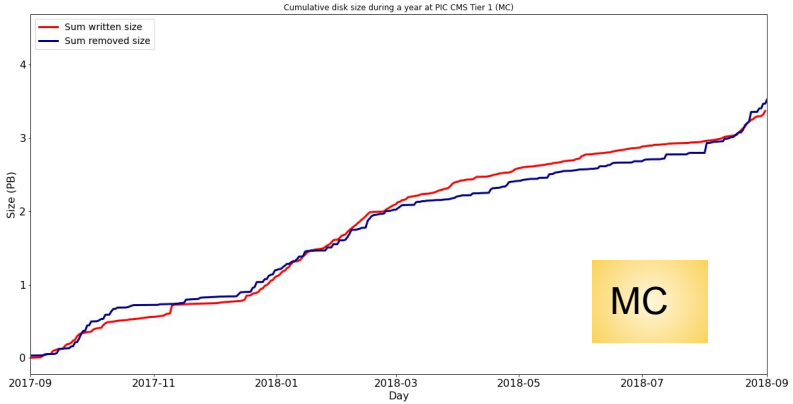
Data popularity from dCache 2/2

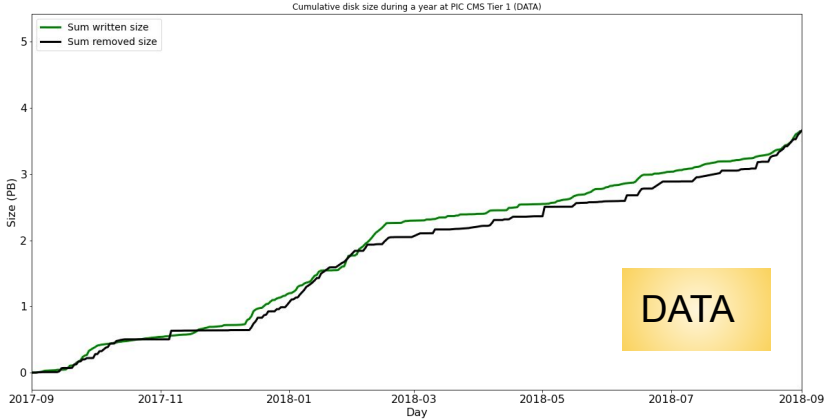
We can also get data popularity per type of files and access protocol



PIC
port d'informació
científica

Write/delete data rates by data tier





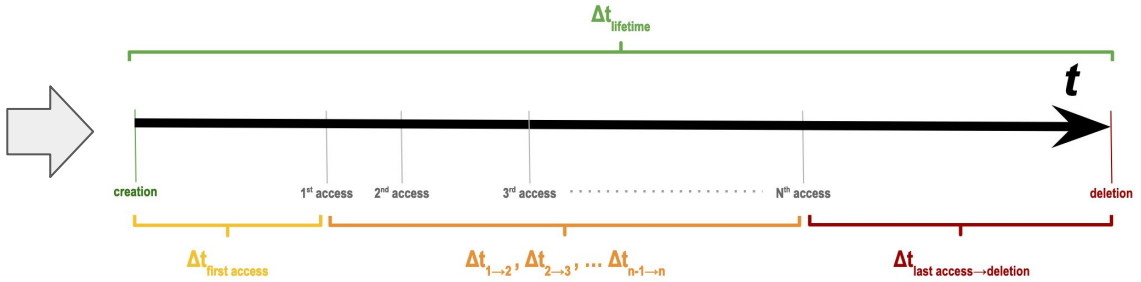
Daily average of writes and removals on disk at PIC is 24 TB/d

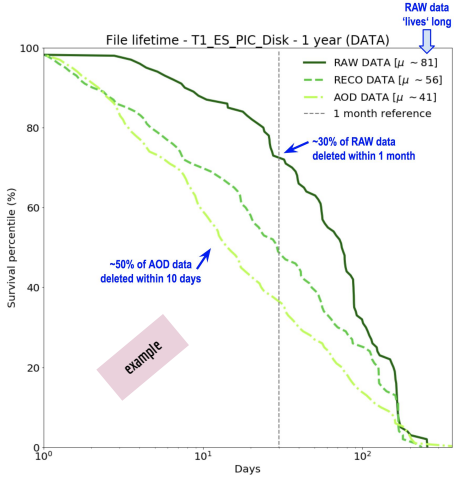
Total disk (TB/d)	MC (TB/d)	DATA (TB/d)	UNMERGED (TB/d)	SAM+LOAD+REST (TB/d)
24	5.7	10.6	3.3	5.4

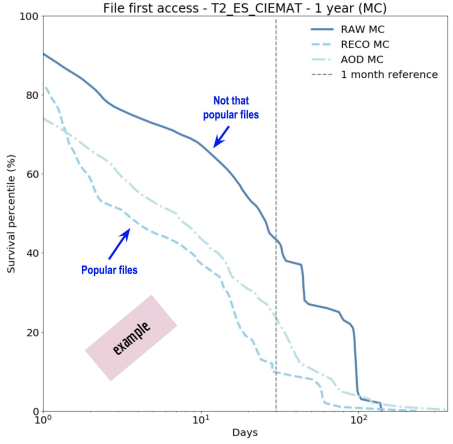
PIC
port d'informació
científica

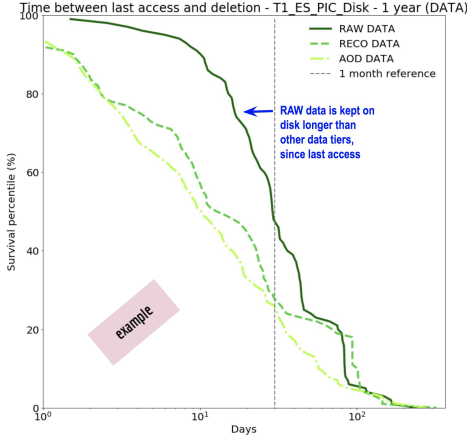
The adventurous life of a file

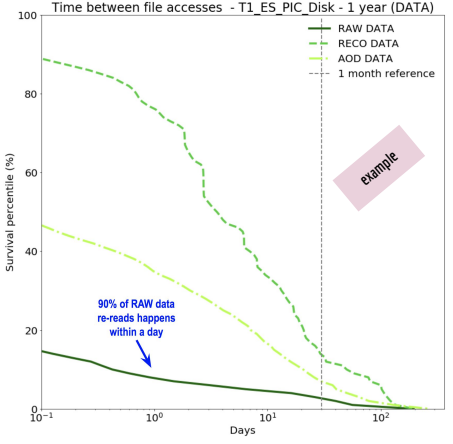
File lifetime
File first access
Time between file accesses
Time from last access to deletion











PIC
port d'informació
científica

Data accesses from CMS jobs monitoring

Another interesting view is what jobs actually read when executed - file access based analysis

CMS jobs monitoring is kept at several Monit databases at CERN:

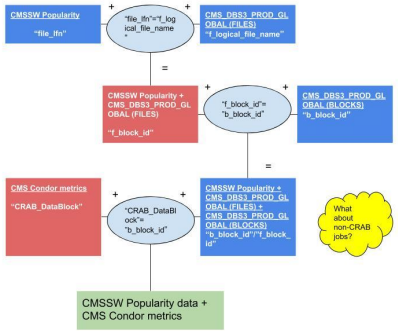
→Information from HTCondor ClassAds

→Information from CMSSW on data accesses (aka popularity)

This data can be analysed at the **Hadoop Spark cluster** at CERN

→Involves multiple joins (sometimes lacking of unique_id to correlate DDBB)

→Not all of the relevant information is available for all of the jobs (accesses at block level beyond CRAB, specific accessed lfn… or CRAB vs production)

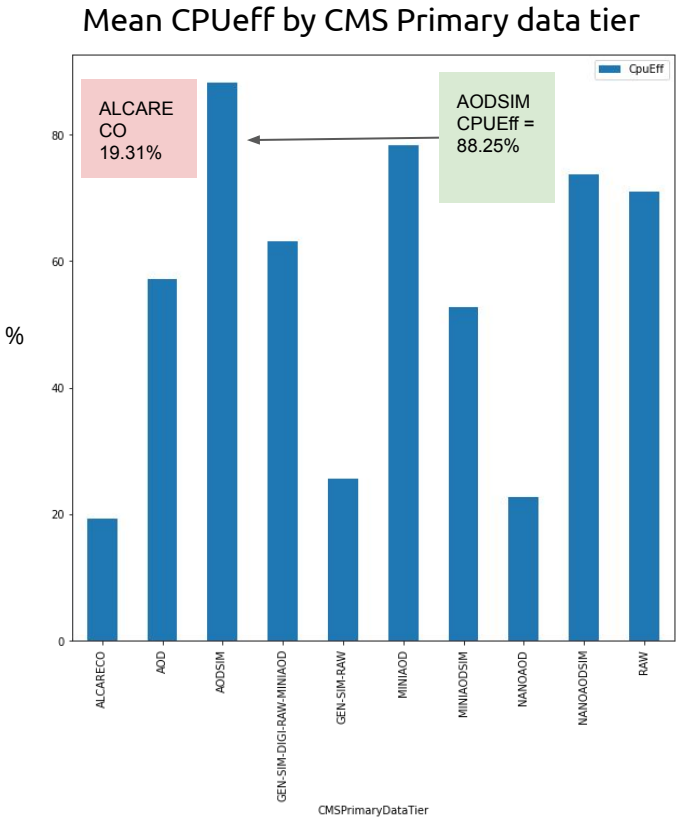


Preliminary studies performed

PIC
port d'informació
científica

CPU efficiency for CRAB jobs

We can extract data from jobs to compute the average CPU efficiency by data tier and identify which CRAB jobs have the best performance



PIC
port d'informació
científica

Reads/Writes to PIC xCache

PIC xCache is serving data to one compute node at PIC since March 2020 (cache-all - minimal setup)

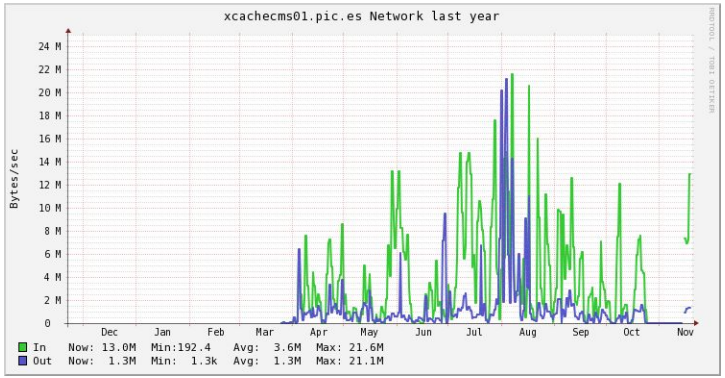
→ As of today the cache is ~55% full

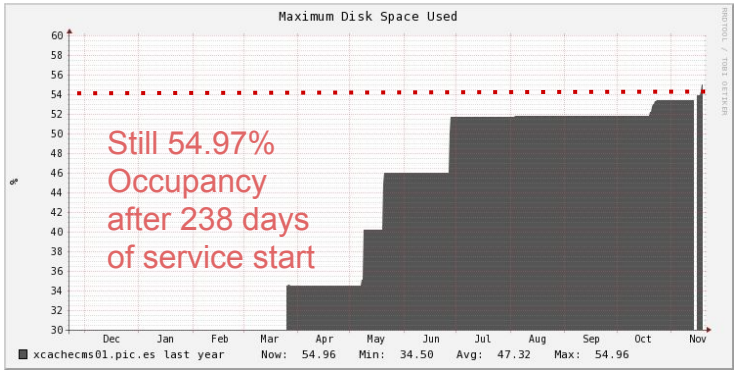
Lacking local xCache monitoring, but ganglia

→ improve when moving to XRootD-5

Relying on CMS job monitoring

→ Dual-stack compute nodes were filtered and file accesses from CMSSW were not seen in the period (recently fixed)
→ Only CRAB jobs were tracked at block level

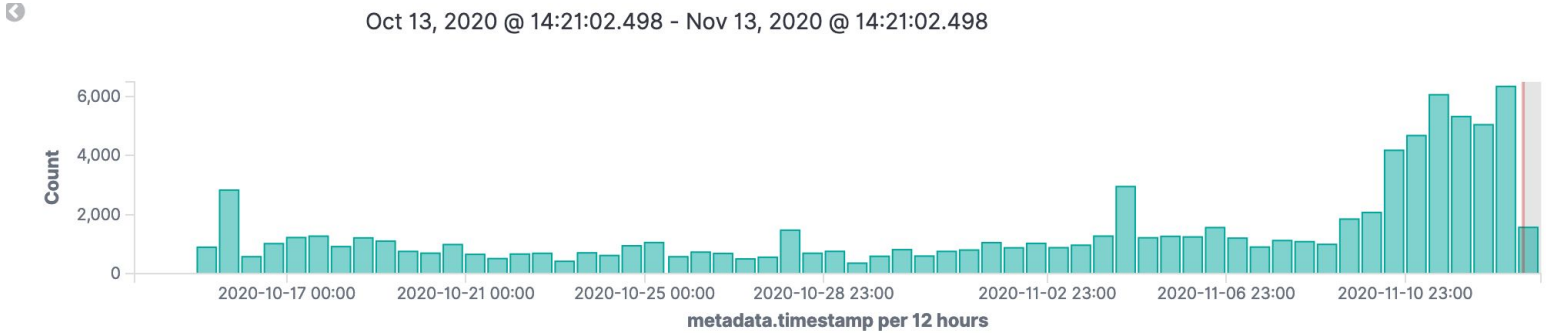




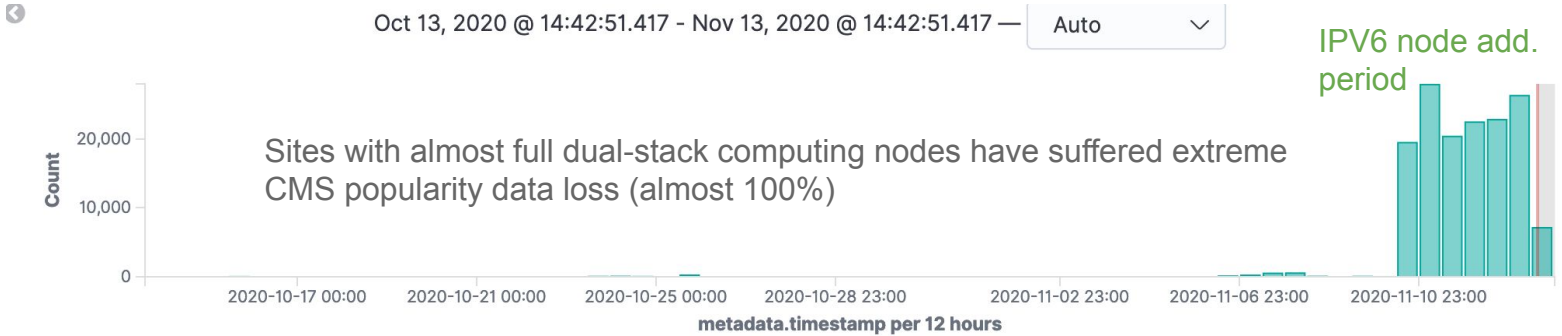
PIC
port d'informació
científica

CMS popularity information

Dual-stack compute nodes were not appearing in the MONIT CMS data popularity tab



PIC CMS T1 (ES)



KIT CMS T1 (DE)

PIC
port d'informació
científica

Looking forward

- **If we want to improve the way in which the storage is offered to WLCG, we need to understand how the storage resources are used**

- **We are approaching this at several levels, for the CMS experiment at PIC:**
  - From the local storage view
  - From the submitted jobs view
  - Latency effects on CMS Workflows

- **xCache instances deployed for CMS in PIC Tier-1 and CIEMAT Tier-2 (Spanish CMS sites)**
  - Expecting to enhance their functionality
  - Working to enable a local monitor that helps improving the service

- **Aiming to get directions to improve the way storage is provided from the region to WLCG experiments**

Thank you for your attention

Questions?