

HEPiX



PIC Report - J. Flix

[on behalf of PIC team]

HEPiX Spring 2023

Taipei (TW)

27-31 Mar 2023



Ciemat



RED ESPAÑOLA DE SUPERCOMPUTACIÓN

España | digital

Unió Europea
Fons Europeu
Next Generation

GOBIERNO DE ESPAÑA

Plan de Recuperación,
Transformación
y Resiliencia

Next Generation
Catalunya

Generalitat de Catalunya
Departament de Recerca
i Universitats

GOBIERNO DE ESPAÑA

SECRETARÍA DE ESTADO DE POLÍTICA ECONÓMICA Y TRANSFORMACIÓN DIGITAL

SECRETARÍA DE ESTADO DE INNOVACIÓN Y TRANSFORMACIÓN DIGITAL

PIC in numbers

March 2023

CPU: 114 kHS06
Disk: 18.3 PB
Tape: 69.5 PB



Spanish WLCG Tier-1 centre → ~80% of resources

→ Provides ~4% of Tier1 data processing of CERN's LHC detectors ATLAS, CMS and LHCb

¼ of the Spanish ATLAS Tier-2 and **a Tier-3 ATLAS data analysis facility** → ~10% of resources

T2K [neutrinos], **MAGIC** and **CTA** [gamma-ray astronomy], **PAU** and **EUCLID** [cosmology], **VIP** [instrumentation], opportunistic access to **LIGO/VIRGO** and **DUNE**, among others...

137 compute nodes (9000 slots), under **HTCondor v.9.0.17**

- Very old hardware switched off this winter (10% CPU reduction)
- New AMD EPYC 7452 32-core processors purchase (1024 vcores - 12.3 HS06/vcore)
- x20 additional servers to come (AMD EPYC 7502, 2048 vcores)
- 100% of compute nodes in dual-stack

2x HTCondor-CE v5.1.6-1.e17

2x ARC-CE v.6.17.0-1 (used by ATLAS and LHCb as HPC gateways - see later)

HTCondor setup in PIC is compatible with SciTokens. Both ATLAS and CMS submitting jobs to our HTCondor-CEs using tokens

18 GPUs available: via JupyterHub and (direct) acces by some VOs, also available through Grid

- 8 GeForce RTX 2080 Ti, 8 Tesla V100-SXM2-32GB, 2 GeForce GTX 1050 Ti

PIC disk storage

~18 PB running on **dCache 8.2**

- New pools acquired to replace obsolete pools and increase capacity:
 - * 15x SuperStorage SSG-6028R-E1CR24N: 24 HDD SAS*18TB (~360 TB neto) and 2x25Gbps NIC
- dCache pools in dual-stack
- TPC enabled for HTTPs and XRootD and token authentication (PIC in DOMA testbeds)
- dCache DDBB upgraded to PostgreSQL14

StashCache deployed as docker container (OSG repo) for Virgo/Ligo

- 3.2 TB - 95% occupancy
- Running XRootD 5.4.2 (OSG 3.6)

xCache deployed (OSG repo) for the CMS experiment

- 6TB disks (RAID60-175 TB). 48 cores E5-2650L v3 (HT enabled). 128 GB RAM. Bonding active-active 10 Gbps - 90% occupancy
- Running XRootD 5.5.1
- Caching *AOD* files off-site, also acting as XCache for the Spanish CIEMAT Tier-2 site

Expansion of the new tape library

IBM TS4500



NEW

IBM TS4500 (64 PB capacity):

- 5 frames (L55+D55 + 3xS55) + 10 LT08 drives + 11 LT09 drives
- 4.8 PB capacity installed with cartridges LT07 M8
- 36.4 PB capacity installed with cartridges LT08
 - * New 1850 LT08 tapes arrived → 59.5 PB in LT08
- Another S55 frame and a second robotic arm to be purchased, which will increase system redundancy

SOON TO BE
RETIRED
THE COUNTDOWN BEGINS...

SL8500



This library is growing to host future data

- It hosts new data and data migrated from SL8500 library (**finished**)
- Dedicated drives, frames and cartridges installed to handle this

All new data writes go to the IBM (LT08)

PIC currently runs **Enstore 6.3.4-14** (CentOS 7)

- We started testing CTA as a potential replacement

Enstore to CTA (initial tests)

dCache 8.2 and **CTA 5.7.12-2**, with the CTA rpms distributed by CTA team

We have a **working instance** that writes and reads correctly. Split IBM TS4500 tape library with a small logical library for testing with **two tapeservers, two drives and some tapes**

Started to **send metrics** to our monitoring system, by tapepool, by queues and by mediatype. We just built small bash scripts using CTA JSON output to send it to Graphite, and plot it with Grafana

DESY's dCache and **FNAL's** Enstore teams are helping us a lot, since they are both testing CTA with a very similar ecosystem than us – **Thank you!**

Also **thanks to the CTA community**. We wanted to show our interest but also ask for help with errors/problems we may have or find. Very useful and active community

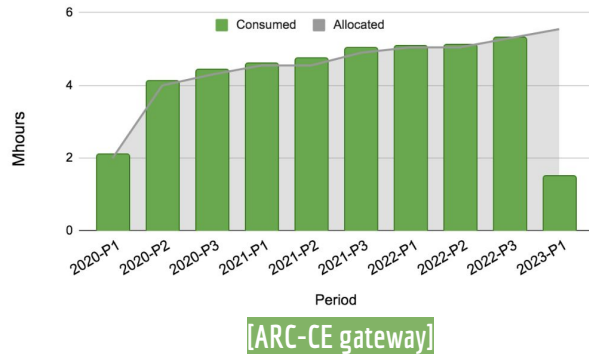
Use of the BSC resources

The LHC experiments have utilized **83 million hours** of resources at the Barcelona Supercomputer Center (BSC) MareNostrum4 HPC facility through services installed at PIC **since 2020**

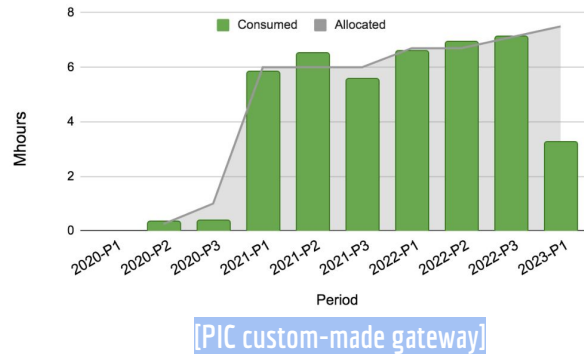
This corresponds to an average installed capacity of approximately 53 kHS06, representing around **30% of the current grid resources deployed at PIC** for the LHC experiments

The new period, 2023-P1, started in March 1st 2023

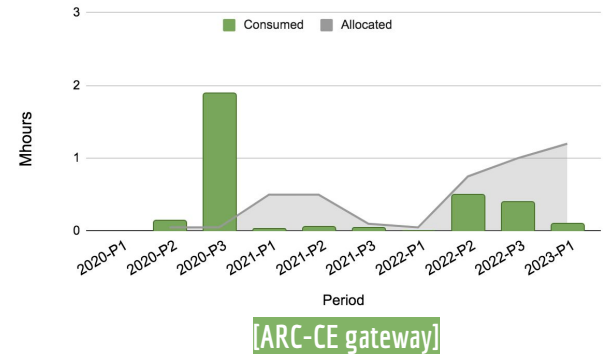
ATLAS



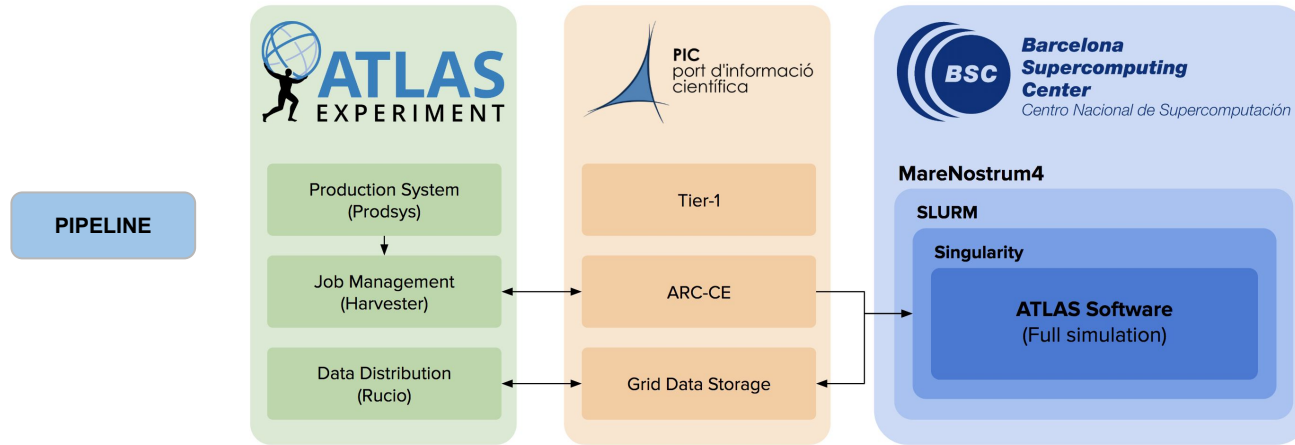
CMS



LHCb



Use of the BSC by ATLAS PIC Tier-1



Submitting **ATLAS** payloads to BSC from PIC Tier-1 since 2018, in production since 2019

Using two **ARC-CEs** at PIC to interconnect MareNostrum and ATLAS production system

Only simulation workflow validated - singularity containers, pre-placed at MareNostrum GPFS

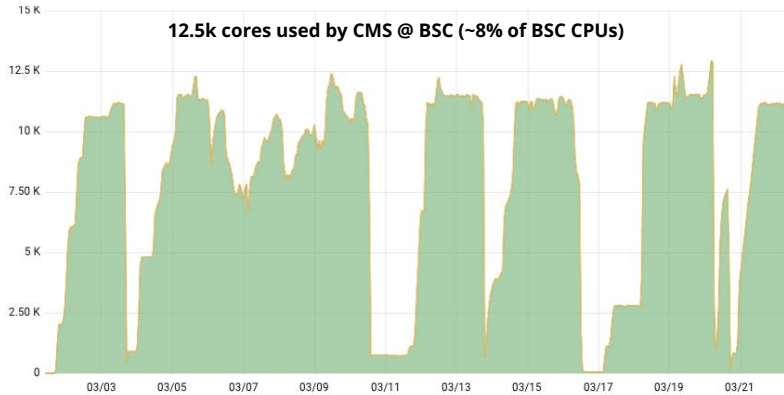
~**15 million hours** approved and used at BSC **every year** by ATLAS through PIC gateways

Other gateways available at the Spanish ATLAS Tier-2s

→ At CHEP2021 proceedings ([link](#))

Use of the BSC by CMS PIC Tier-1

Running cores



CVMFS CMSSW monitoring



Current status

- Running in operations (**CMS Simulation workflows GEN-SIM**)
- Result of the **PIC and HTCondor team collaboration** to use a **shared FS as control path for HTCondor**
- Interaction with BSC execute nodes through the login node, mounting the shared FS through **sshfs** and sending jobs to the Slurm scheduler via **ssh**. Slurm jobs launch a HTCondor slot that joins the **CMS Global Pool**
- **CMS Software modified** to accept sql files for conditions data at runtime
- Using **cvmfs_preload** to bring cvmfs CMS files to BSC. Two weeks to copy ~37M files (13 TB), at first injection. **cvmfsexec** used to build the cvmfs file structure
- **Stage-in/out + Data Transfer Manager** designed to transfer input and output data from/to PIC (*xRootD server in singularity images*)
- Integrated with **WMAgent @ CERN - Accounting to APEL** ongoing
- A New grant of **7.5 Mhours granted** (typical quarter allocation for CMS)

- At **CHEP2021** proceedings ([link](#))
- At **ISGC 2022** ([link](#))
- At **CHEP2023** ([link](#))

Use of the BSC by LHCb PIC Tier-1

LHCb used similar technical implementations as ATLAS (**ARC-CE02.PIC.ES**) to exploit BSC resources - submitting grants to BSC as ATLAS and CMS, and **modified DIRAC** for the purpose

LHCb		Submit Host	Jul 2022	Aug 2022	Sep 2022	Oct 2022	Nov 2022	Dec 2022	Jan 2023	Feb 2023	Mar 2023	Total	Percent
	es13.pic.es-9619/es13.pic.es-cndr		8,399	11,125	14,649	14,744	14,141	12,499	24,009	14,431	7,968	121,965	34.34%
	es14.pic.es-9619/es14.pic.es-cndr		8,251	11,252	14,314	13,371	14,006	12,116	24,172	13,654	10,359	121,494	34.21%
	gatlh://arc-ce02.pic.es-2811/jobs		1	0	3	0	0	0	0	0	0	4	0%
	https://arc-ce02.pic.es-8443/arcx		13,840	16,418	8,308	16,836	463	15,714	21,962	9,381	8,929	111,657	31.44%
	Total		30,291	38,795	37,274	44,951	28,610	40,323	70,143	37,472	27,255	355,120	
	Percent		8.53%	10.32%	10.50%	12.68%	8.08%	11.30%	19.75%	10.55%	7.67%		

Dask integration

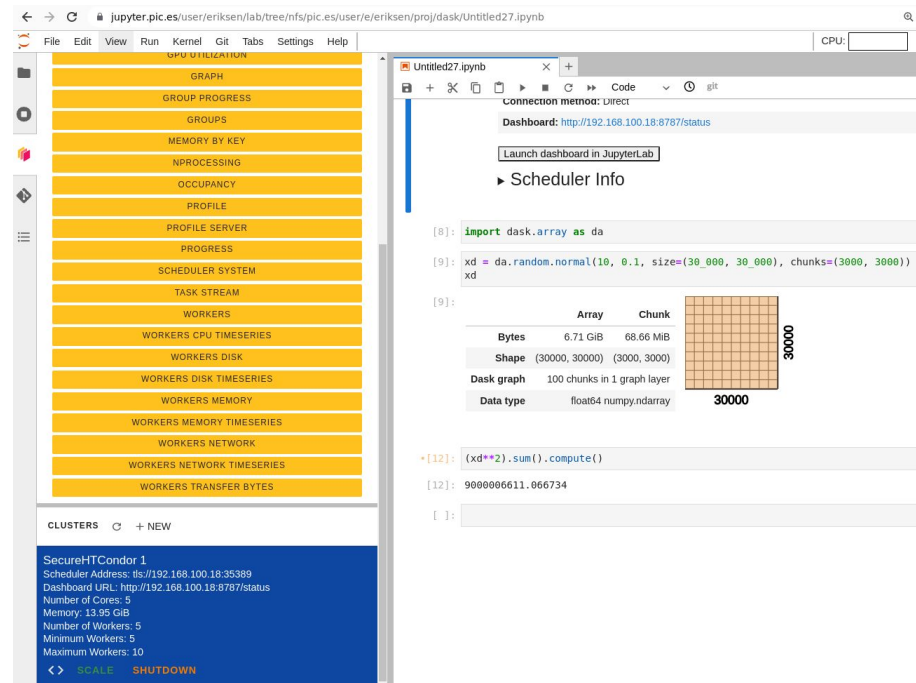
Dask scales data science libraries like Numpy and Pandas to multiple machines

Based on **low level task parallelism**, allowing parallelization of custom codes

Integrated into Jupyter so the user can start and monitor a cluster from the GUI

The started cluster request PIC resources **through HTCondor**

The **adaptive cluster size** can scale up and down based on the workload



The screenshot shows the JupyterLab interface with a Dask cluster monitoring panel on the left and a code editor on the right.

Cluster Monitoring Panel (Left):

- GPU UTILIZATION
- GRAPH
- GROUP PROGRESS
- GROUPS
- MEMORY BY KEY
- NPROCESSING
- OCCUPANCY
- PROFILE
- PROFILE SERVER
- PROGRESS
- SCHEDULER SYSTEM
- TASK STREAM
- WORKERS
- WORKERS CPU TIMESERIES
- WORKERS DISK
- WORKERS DISK TIMESERIES
- WORKERS MEMORY
- WORKERS MEMORY TIMESERIES
- WORKERS NETWORK
- WORKERS NETWORK TIMESERIES
- WORKERS TRANSFER BYTES

CLUSTERS Panel (Bottom Left):

- SecureHTCondor 1
- Scheduler Address: `ts://192.168.100.18:35389`
- Dashboard URL: `http://192.168.100.18:8787/status`
- Number of Cores: 5
- Memory: 13.95 GiB
- Number of Workers: 5
- Minimum Workers: 5
- Maximum Workers: 10
- Buttons: <> SCALE SHUTDOWN

Code Editor (Right):

```

import dask.array as da

xd = da.random.normal(10, 0.1, size=(30_000, 30_000), chunks=(3000, 3000))
xd

[9]:
      Array      Chunk
Bytes      6.71 GiB    68.66 MiB
Shape (30000, 30000) (3000, 3000)
Dask graph 100 chunks in 1 graph layer
Data type  float64 numpy.ndarray

+ [12]: (xd**2).sum().compute()
[12]: 9000006611.066734
  
```

The code execution output includes a visualization of the Dask array's shape and data type, and the result of a summation operation.

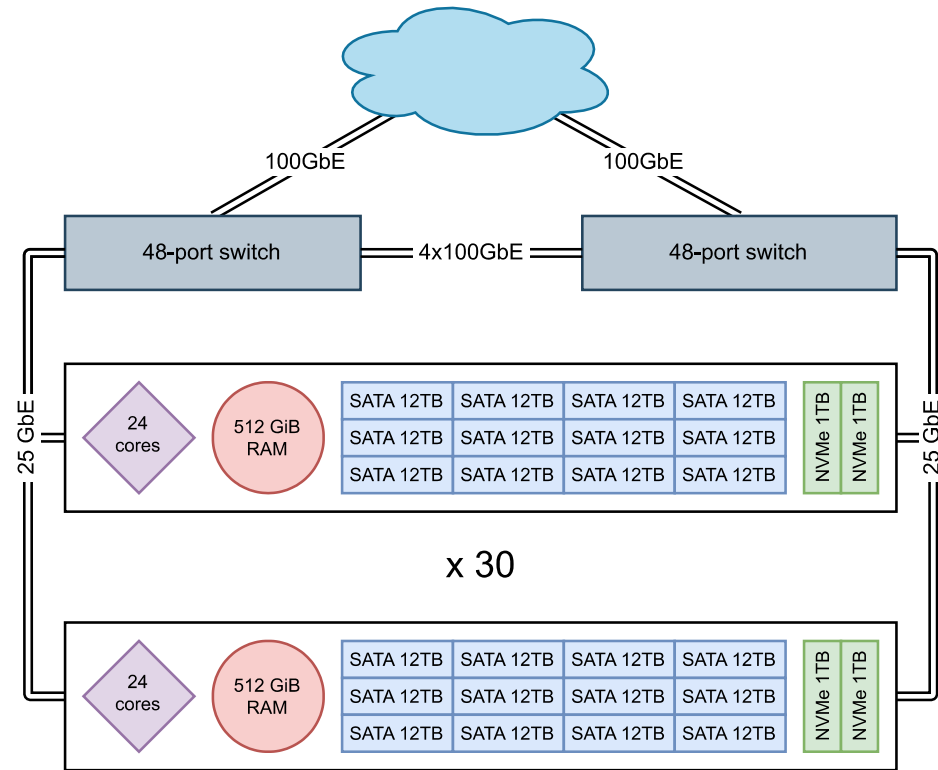
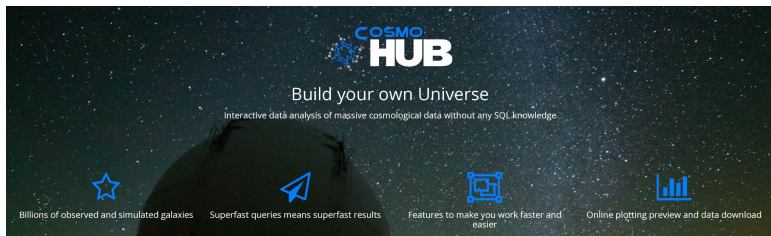
New Hadoop Cluster

30 nodes:

- 720 cores, 15 TiB RAM
- 60 TB NVMe (for cache)
- 4.3 PB raw storage (2.5 PB usable)

Main use cases:

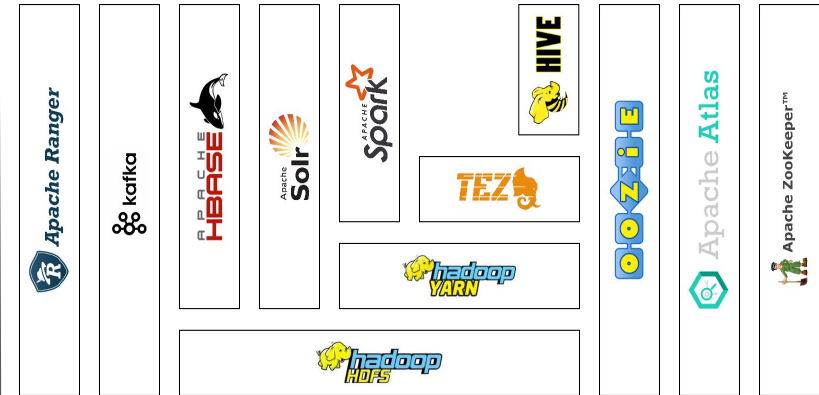
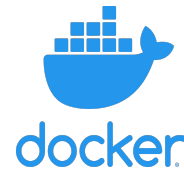
- **CosmoHub** query processing
- Euclid **mock galaxy catalogs**
- **HTCondor backfilling**, specially suited for **ephemeral/adaptive Dask** clusters



Custom Hadoop distribution

Based on a single Docker image:

- Atlas 2.2.0
- **Hadoop 3.2.3**
- HBase 2.2.6
- **Hive 3.1.2**
- Kafka 2.5.0
- Oozie 5.2.1
- Ranger 2.1.0
- Solr 6.5.1
- **Spark 3.1.2**
- Tez 0.10.0
- ZooKeeper 3.7.1



Built, tested and deployed automatically
using our GitLab CI/CD

New applied AI group

New scientific group at PIC from autumn 2022

Works on **deep learning** in different fields, aiming to **developing synergies**

Ongoing work in **cosmology, material science, bio imaging** and **quantum computing**

Collaboration or interactions with **theory, GW and neutrino groups**

Teaching of deep learning methods

Involved in **developing infrastructure**, like the **Dask** integration



InCAEM Project

In-situ Correlative Facility for Advanced Energy Materials

Correlative in situ experiments **combining** (S)TEM (Scanning Transmission Electron Microscopy), AFM/STM (Atomic Force Microscopy / Scanning Tunneling Microscopy) and synchrotron radiation

Structure ↔ Function

Operando & in situ

Multi-modal & multi-lengthscale analysis

Advanced data analysis: HPC/HTC, deep learning,...

PIC will collaborate with ALBA Synchrotron to build and provide the computing infrastructure for data handling and analysis



<https://www.icmab.cat/incaem-workshop-at-alba-synchrotron-on-advanced-materials-imaging>

Quantum Spain Project

PIC participates in Quantum Spain project to **deploy a quantum computer in Spain**. Part of the future user support team

Promoted by the Ministry of Economy through the Secretary of State for Digitization and Artificial Intelligence and financed with the Recovery Funds

Budget: €22 million

Execution: 01/01/22 – 31/12/25

Goals

- Acquisition and installation of a quantum computer based on superconducting qubits technology
- Create a remote access system in the cloud
- Develop useful quantum algorithms, applicable to real problems

<https://quantumspain-project.es/en>



España | digital ²⁰/₂₆



R Plan de Recuperación,
Transformación
y Resiliencia



Financiado por
la Unión Europea
NextGenerationEU



Summary

LHC computing is included in the **BSC strategic projects portfolio**, which allows us (PIC) to use a fraction of their CPU resources for ATLAS, CMS and LHCb experiments

PIC is part of the national **Data services RES nodes**, which also allows us to exploit storage resources as well for the LHC experiments (though this is still marginal)

Lot of **work done at PIC to get prepared for future**, which included a major migration to a new tape storage system and a upgrade/re-design of the WAN connectivity

Funding landscape in Spain is **in better shape now**, as compared to recent years. We hope to profit in the next project calls

PIC center getting **reinforced** in the support for **other scientific disciplines**. PIC selected as **one of the four CTA datacenters, cooperation with ALBA, ...**



Thanks!
Questions?

Credits to: E. Acción, V. Acin, C. Acosta, A. Alou, M. Børstad, A. Bruzzese, L. Cabayol, E. Carrasco, J. Carretero, J. Casals, R. Cruz, M. Delfino, J. Delgado, J. Flix, E. J. González, D. Graña, G. Merino, C. Neissner, A. Pacheco, C. Pérez, A. Pérez-Calero, E. Planas, M.C. Porto, B. Rodríguez, P. Tallada, J. Priego, F. Torradeflot

www.pic.es

Acknowledgements

The authors of this work express their gratitude to the PIC and CIEMAT teams for their support in these studies and for deploying novel cache services for the CMS experiment in the Spanish region. This project is partially financed by the Spanish Ministry of Science and Innovation (MINECO) through grants FPA2016-80994-C2-1-R, PID2019-110942RB-C22 and BES-2017-082665, which include FEDER funds from the European Union. It has also been supported by the Ministerio de Ciencia e Innovación MCIN AEI/10.13039/501100011033 under contract PID2020-113614RB-C21, the Catalan government under contract 2021 SGR 00574, and the Red Española de Supercomputación (RES) through the grant DATA-2020-1-0039.



Thanks!
Questions?

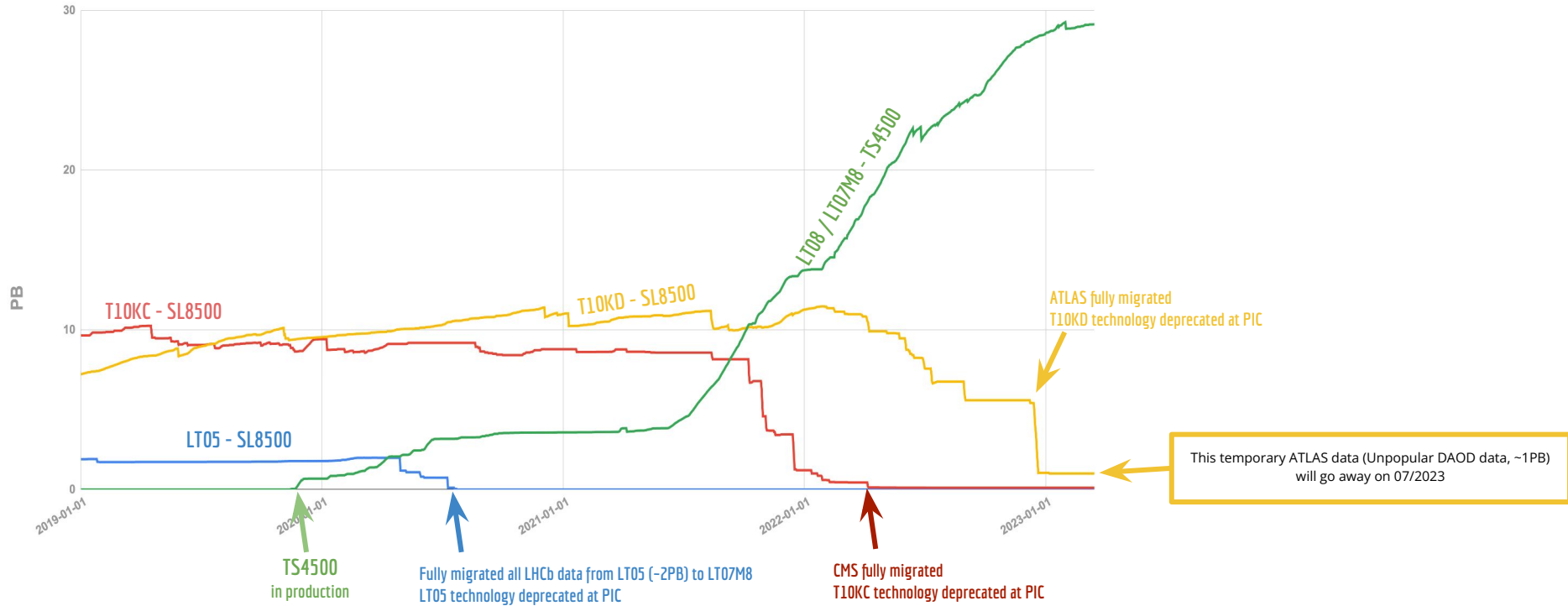
Credits to: E. Acción, V. Acin, C. Acosta, A. Alou, M. Børstad, A. Bruzzese, L. Cabayol, E. Carrasco, J. Carretero, J. Casals, R. Cruz, M. Delfino, J. Delgado, J. Flix, E. J. González, D. Graña, G. Merino, C. Neissner, A. Pacheco, C. Pérez, A. Pérez-Calero, E. Planas, M.C. Porto, B. Rodríguez, P. Tallada, J. Priego, F. Torradeflot

www.pic.es

Backup slides

Data Migrations to TS4500

Used space by WLCG

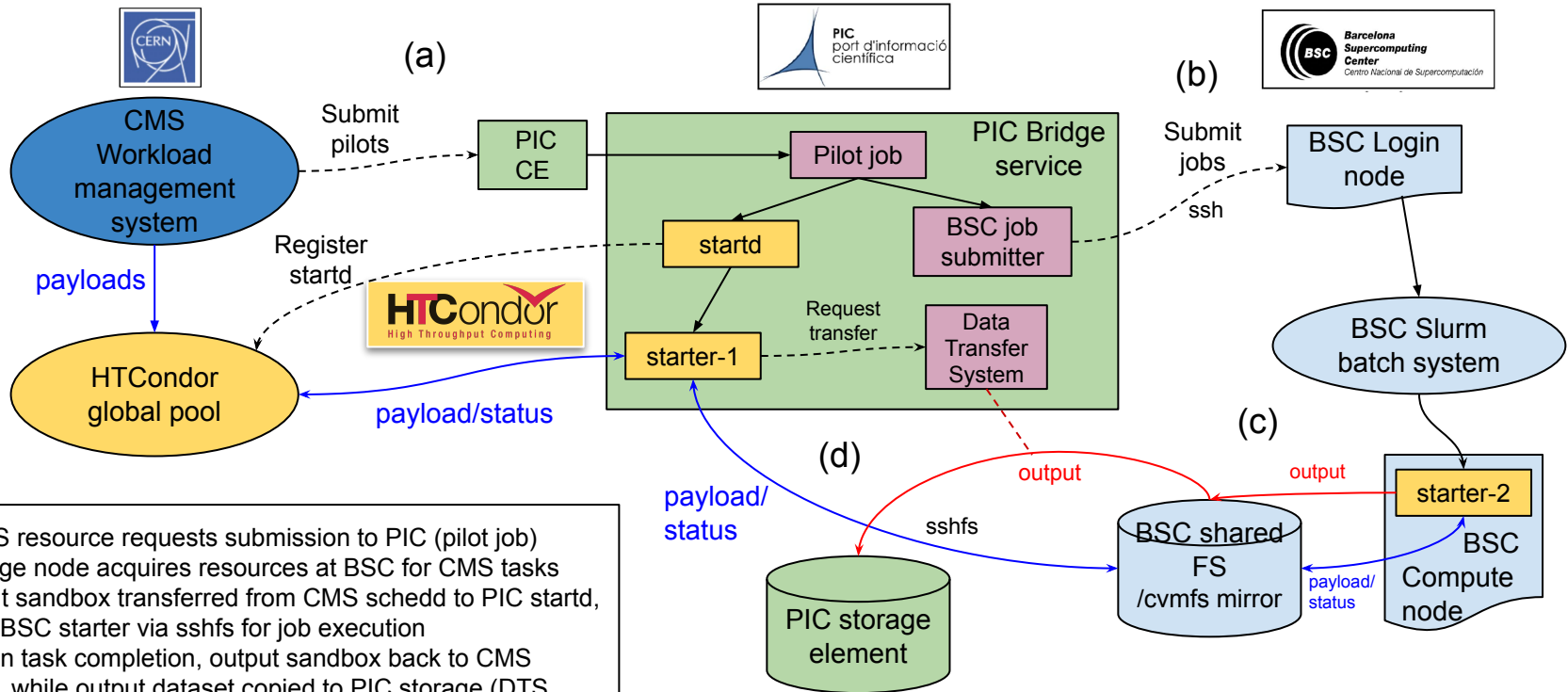


CentOS 7 EOL

PIC participated and agreed on the proposal resulting from the **GDB discussions**

We started deploying **Rocky Linux 8** as a substitute for CentOS7, but given the recent announcement from CERN and FNAL, and the discussions in the Linux Future forum, we are thinking to **move to Alma Linux 9 at some point in future**

Use of the BSC by CMS PIC Tier-1

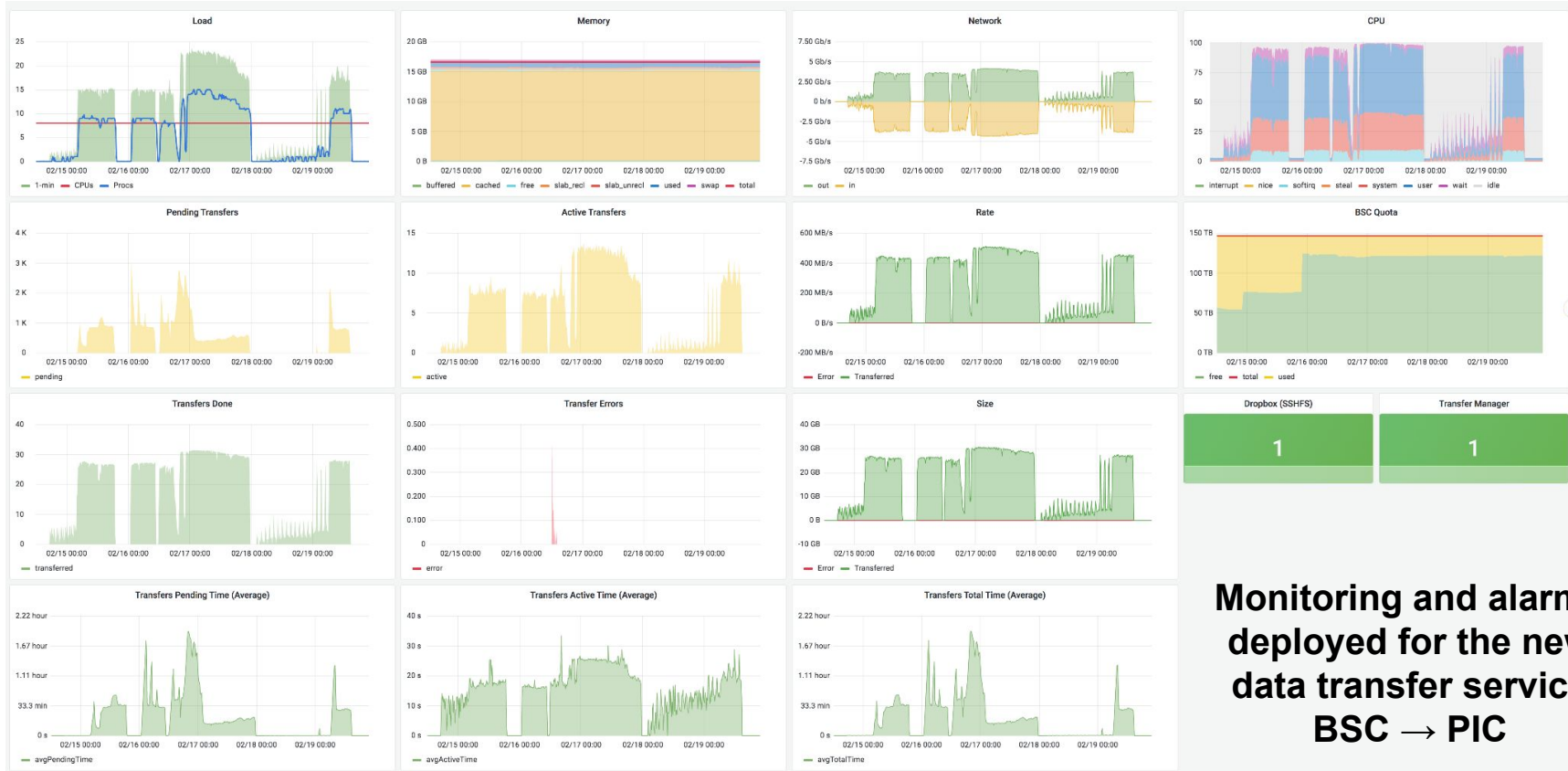


(a) CMS resource requests submission to PIC (pilot job)
 (b) Bridge node acquires resources at BSC for CMS tasks
 (c) Input sandbox transferred from CMS schedd to PIC startd, then to BSC starter via sshfs for job execution
 (d) Upon task completion, output sandbox back to CMS schedd, while output dataset copied to PIC storage (DTS acting as third party copy manager)

PIC and HTCondor team collaboration to use a shared FS as control path for HTCondor

- At CHEP2021 proceedings ([link](#))
- At ISGC 2022 ([link](#))
- At CHEP2023 ([link](#))

Use of the BSC by CMS PIC Tier-1



**Monitoring and alarms
deployed for the new
data transfer service
BSC → PIC**