Integration of BSC resources into CMS Computing

A. Delgado, J. Flix, J.M. Hernández, A. Pérez-Calero, E. Pineda

ISGC 2022, March 24th 2022

GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA E INNOVACIÓN
Ciemat Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas
cfp CIEMAT física de partículas
PIC port d'informació científica

Outline of the talk

- PIC and BSC Context
  - Barriers for resource exploitation by CMS
- Solutions applied:
  - Overcoming network restrictions for job management: HTCondor modifications
  - Bringing code and experimental conditions data to BSC
    - CMSSW and CVMFS
    - Conditions
  - Moving output data from BSC to CMS storage: Data transfer service
- Scalability test results and current status
- Next steps

- PIC AND BSC CONTEXT

Intro to PIC

The Port d'Informació Cientifica, PIC, is the largest Worldwide LHC Computing Grid centre in Spain

PIC supports a Tier-1 site for ATLAS, CMS and LHCb LHC experiments, along with Tier-2 and Tier-3 services for ATLAS

PIC also supports many other data-intensive scientific research fields (Neutrinos, Astroparticle Physics, Cosmology, etc)

Massive storage in tape library (~50 PB installed)

Processing capacity: 10k CPU cores+GPUs. Liquid cooling WNs

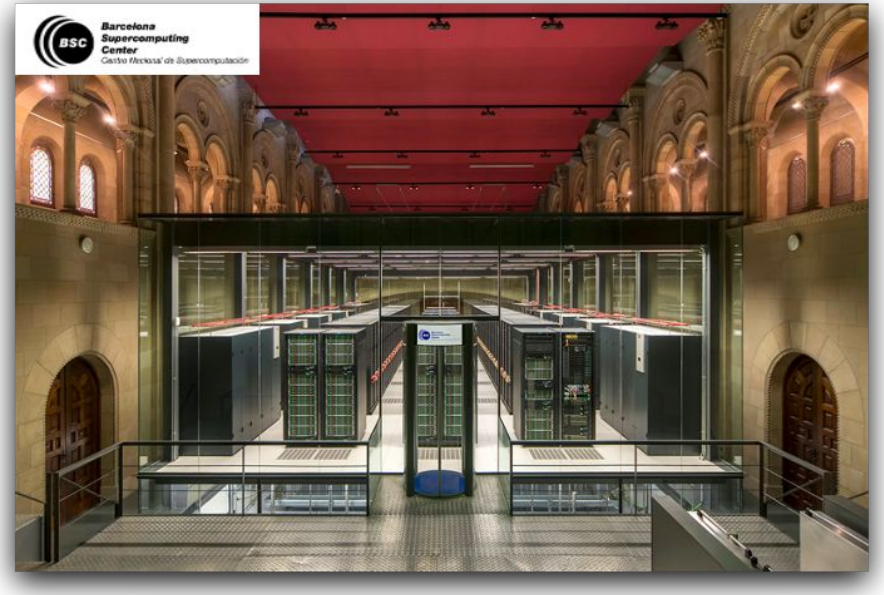
PIC as part of the Spanish Supercomputing Network (RES)

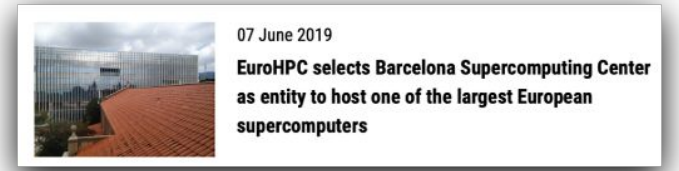
The Barcelona Supercomputing Center

The **Barcelona Supercomputing Center**, BSC, is the main HPC center in Spain

Their main general purpose cluster is named MareNostrum4, with over **150k processors** (**11.15 Petaflops**)

Selected as the site for deployment of a new European Pre-Exascale HPC cluster (MareNostrum5, about **200 Petaflops**)





PIC
port d'informació
científica

The BSC context

- Pressure from Spanish funding agency (Ministry of Science) to use HPC facilities for WLCG computing
  - Flat funding for WLCG resources, lot of public funding went to HPCs

- After lengthy discussions, a **collaboration agreement** was signed with the BSC management, resulting in LHC computing being designated as a strategic project
  - Agreement promoted by WLCG-ES community
  - Guaranteed use of up to **7% of MareNostrum4** (~100M coreHours/year max.)
  - WLCG-ES request: all CPU time required for LHC simulation in Spain
    - ~55 Mhours in 2021 (ATLAS 50%, CMS 30%, LHCb 20%)
  - Submission of proposals for time allocation every 4 months (**fast lane approval process**)

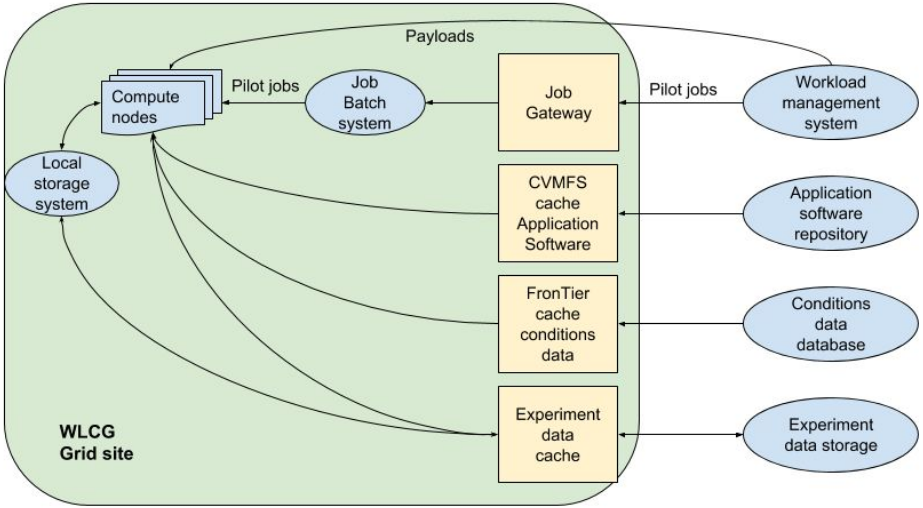
BSC constraints to run CMS jobs

Using BSC's MN4 resources is **very challenging** for CMS

- **No Internet connectivity in compute nodes**!
  - A showstopper for CMS, as tasks require access to external services
  - Experiment edge services not allowed inside BSC

- Available access:
  - A login node which allows ssh
  - A shared disk (GPFS) mounted on execute nodes and login machines - accessible from outside via sshfs

Substantial integration work to make BSC capable to run CMS jobs



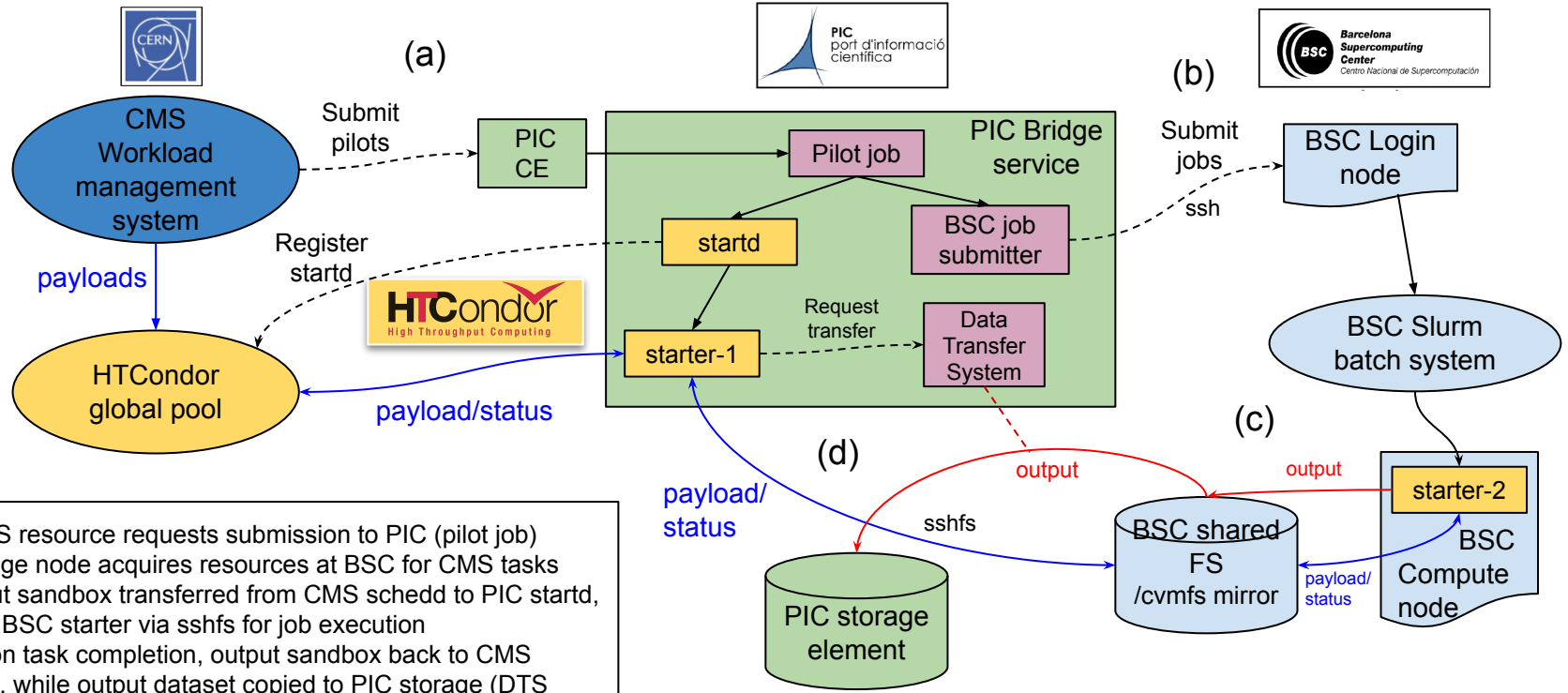
- SOLUTIONS APPLIED

PIC
port d'informació
científica

Solutions applied

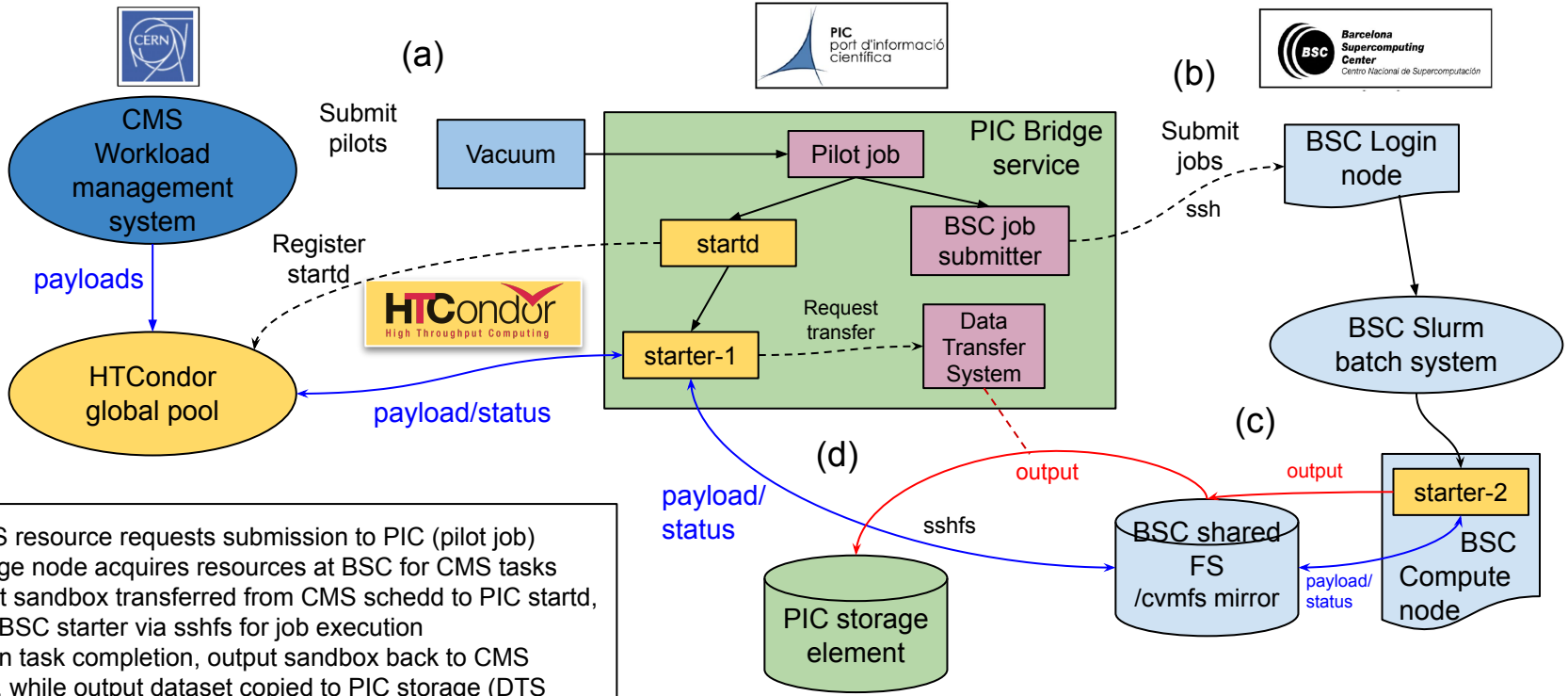
- HTCondor major development: modify CMS resource provisioning, job scheduling and execution framework (HTCondor) to use a **shared file system as communication layer**
  - **Split-starter model,** presented at CHEP19
  - Requires a **bridge service at PIC** to connect CMS WMS and BSC
- CMS software (**CMSSW**) deployed to BSC environment via CVMFS pre-loaded replica
- **Conditions data** read from sqlite files, pre-placed in BSC GPFS
- Local environment configuration for CMS tasks (e.g. where to write output data)
- **Develop** and **operate** a **custom data transfer service** (DTS) for output data migration from BSC to PIC storage

HTCondor split-starter + DTS



(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)

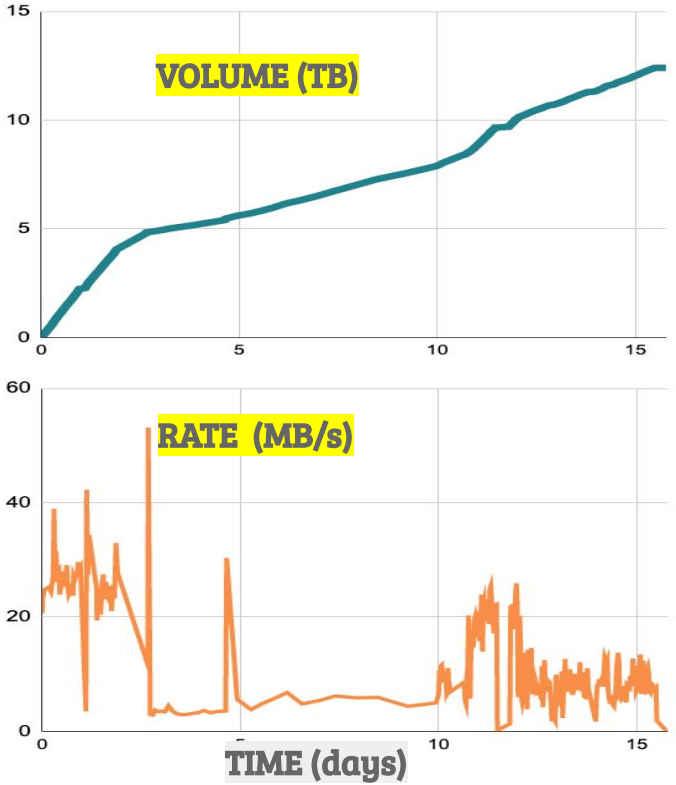
HTCondor split-starter + DTS



(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)

Deploying full CMS CVMFS tree

- Copied whole *cms.cern.ch* repository:
  - 12.6 TB, **183M files** (37M files de-duplicated)
- Used *cvmfs_preload* tool
  - Avoids duplication, skips already present files
  - Run at PIC, directly into an SSHFS mount of BSC shared filesystem
- Took ~2 weeks to complete
  - After initial phase with frequent transfer errors
    - Intervention at stratum 0 was required
  - Directories with large files showed higher rates
- Following periodic updates (*deltas*) is much faster



The Data Transfer Service (I)

A custom Data Transfers Service (DTS) has been designed and implemented in order to manage **output data transfers** from BSC to PIC storage:

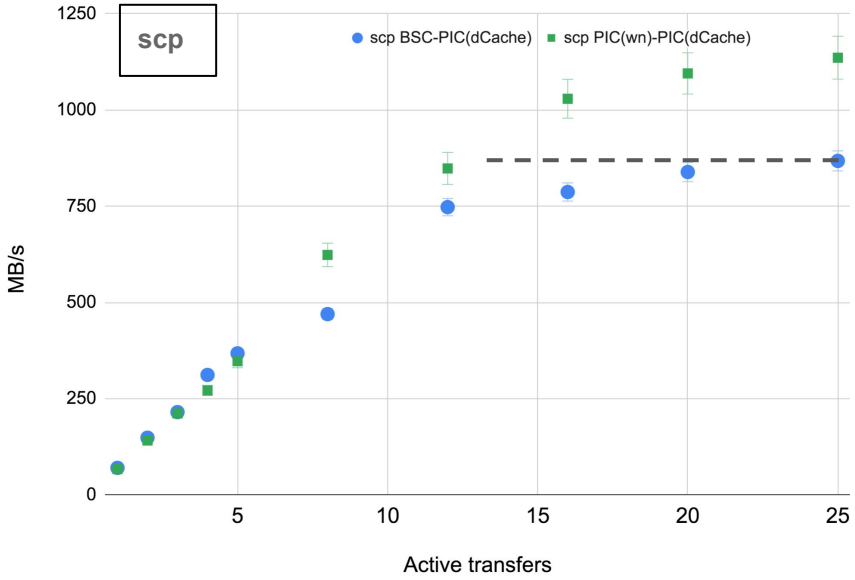
- **Synchronous data transfer:** requests launched by ad-hoc job wrapper, jobs wait O(10) secs until transfer completed
- DTS running on a dedicated node at PIC executes **parallel scp transfers** between BSC and PIC storages as triggered by the ad-hoc job wrappers
- The transfer is **transparent to CMS WM** services, assume data is already at PIC when job finishes, then proceeds to register the produced dataset to PIC_Disk RSE
- Fractional output data files (*unmerged files*) generated at BSC are copied to PIC, then standard CMS *merge* jobs are executed at PIC to produce full size dataset files

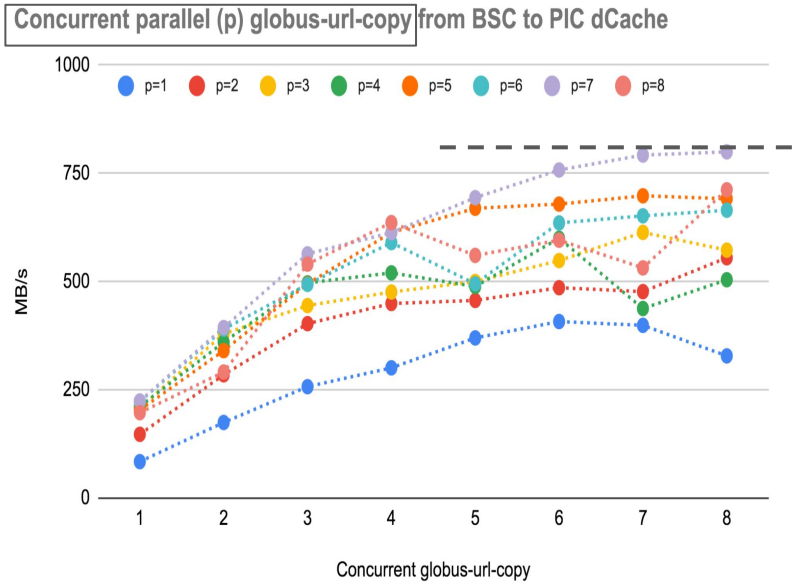
Initially, **focus on workflows with no input** (GEN-SIM). The rest of the chain (DIGI-RECO, I/O intensive and needs to read input data files) can then be executed at PIC (or other sites remotely reading those inputs from PIC).

- Under consideration: **Input data** could potentially be served also by a similar DTS

The Data Transfer Service (II)

- Tested **single stream multi-file-copies** and **multithreaded** methods (scp's, cp's, globus-url-copy).
- Transfers from BSC **saturating** at ~800 MB/s
  - On a **10 Gbps network link**, so about ⅔ of the max theoretical capacity in use
  - **Similar saturation** regardless of the copy mechanism
- For simplicity (e.g. no extra authentication step), **scp** selected as the DTS method to copy files from BSC GPFS to PIC dCache.





The Data Transfer Service (III)

Monitoring and alarms deployed for the new data service

- SCALABILITY TEST RESULTS AND CURRENT STATUS

PIC
port d'informació
científica

Scalability results

- Once the **solutions** described were **deployed**, CMS workflows producing MC simulated datasets were assigned to run at BSC in order to **fully commissioning the new setup**.

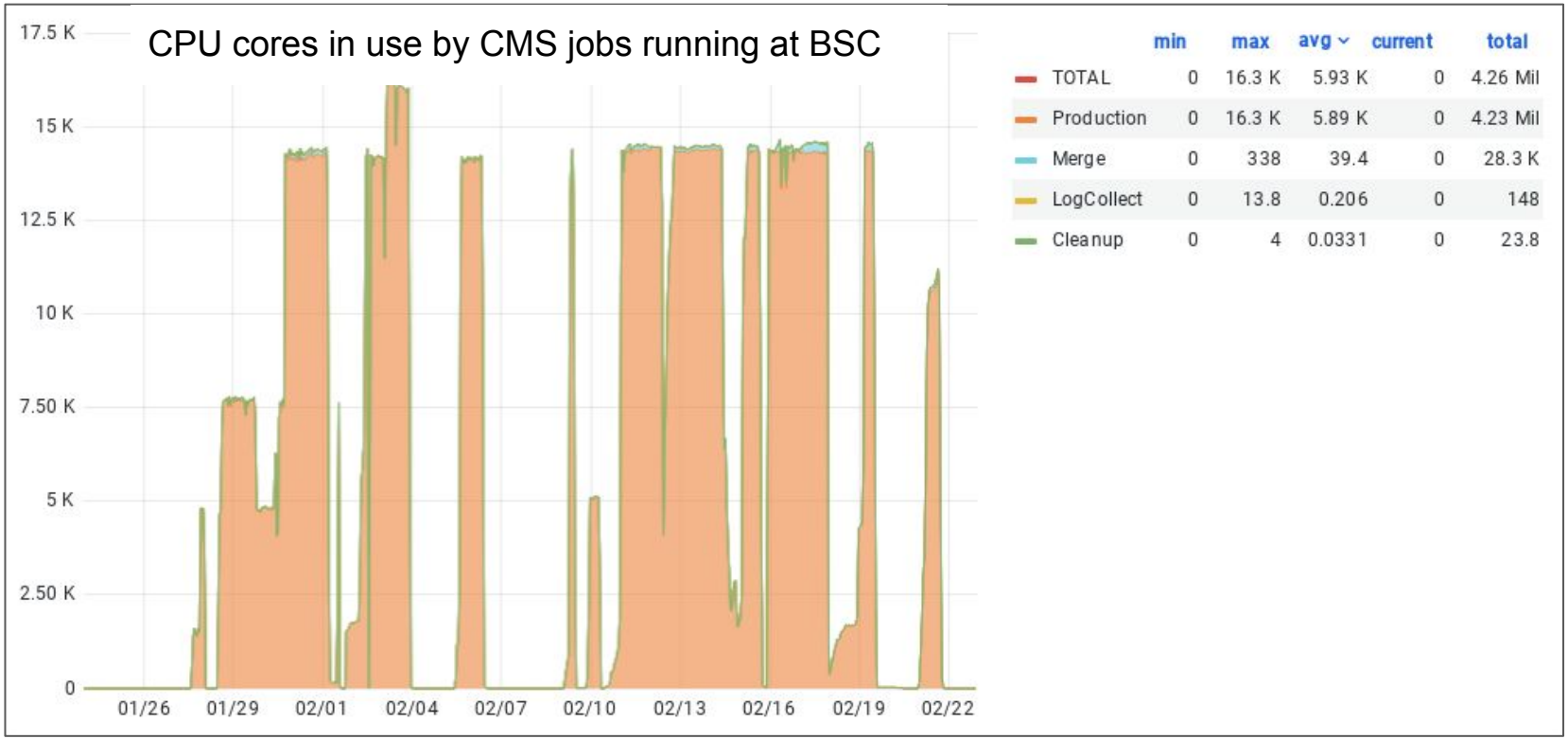
- These test workloads were finally executed at BSC to completion successfully, with **comparable error rate and CPU efficiency** to CMS tasks on regular WLCG slots

- The new DTS correctly handled file transfers into CMS disk storage at PIC, without introducing additional CPU inefficiency

- The infrastructure and services deployed were proven capable to sustain a scale of **~15k CPU cores in BSC's MN4** (**500 MB/s aggregate output rate**)

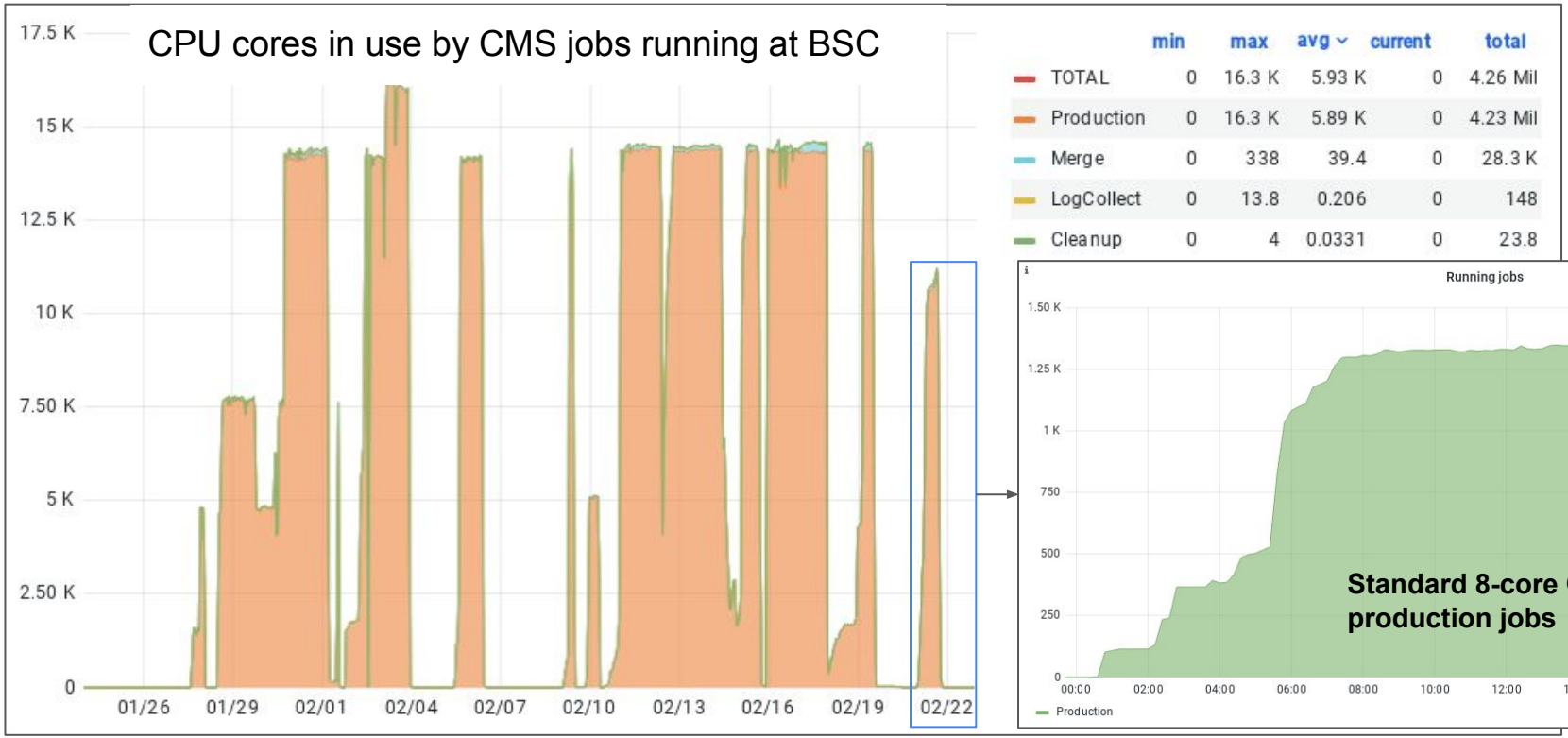
Execution of CMS workloads at BSC

Integration and tests at scale executed during the Jan-Feb of 2022, with successive corrections and improvements



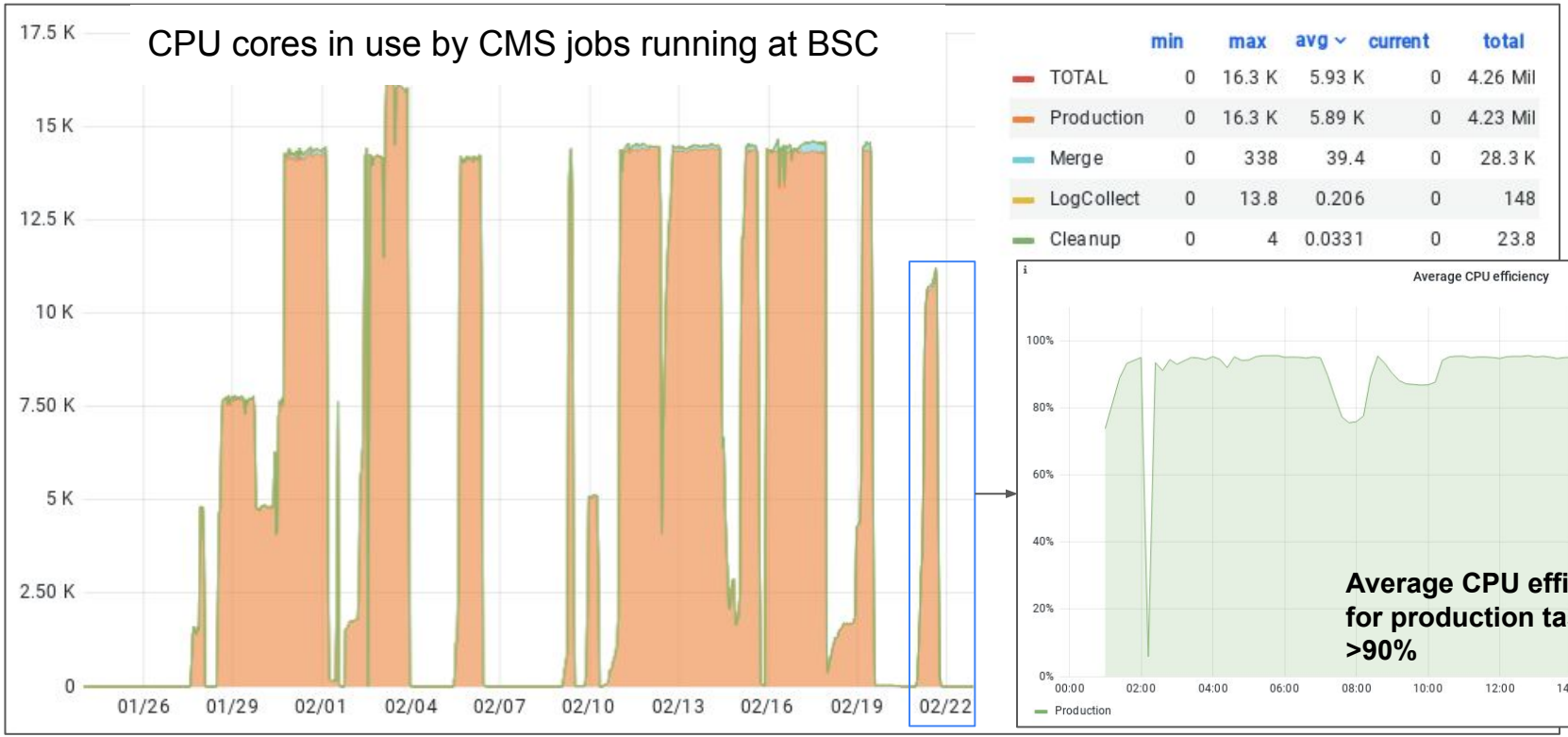
Execution of CMS workloads at BSC

Integration and tests at scale executed during the Jan-Feb of 2022, with successive corrections and improvements



Execution of CMS workloads at BSC

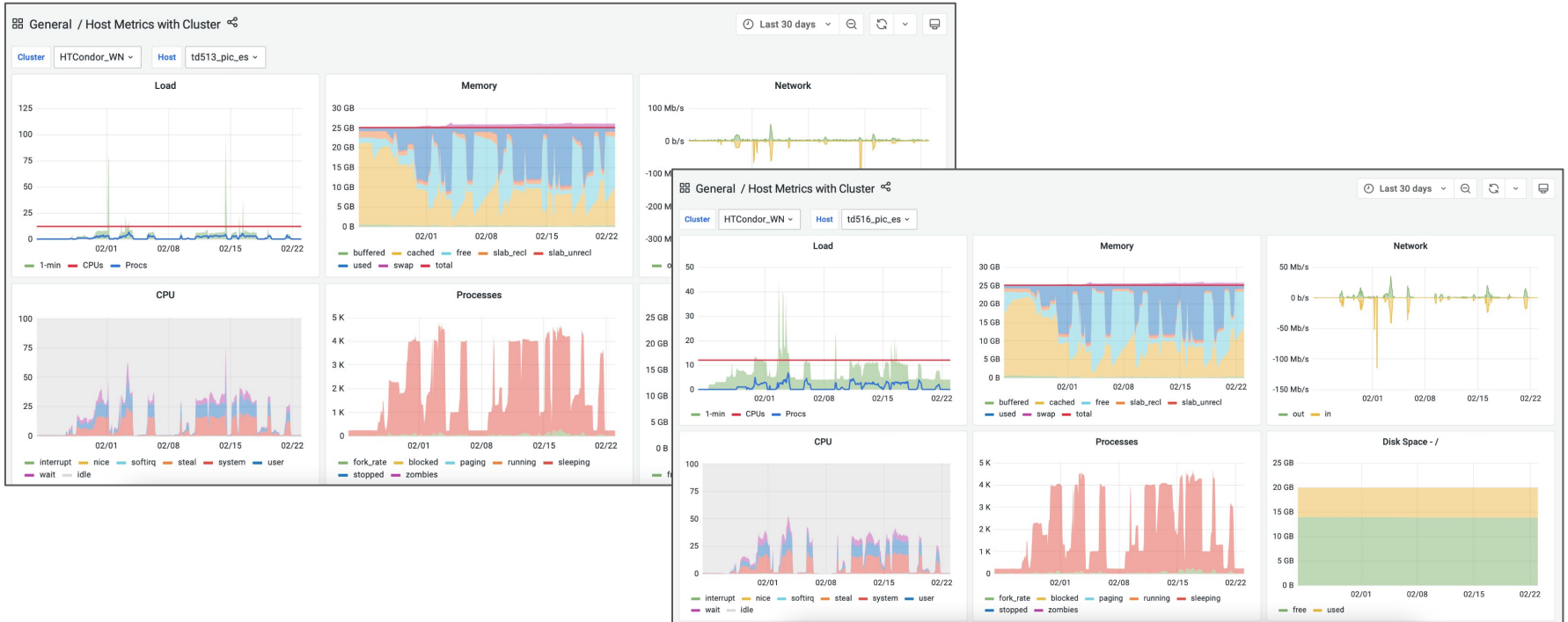
Integration and tests at scale executed during the Jan-Feb of 2022, with successive corrections and improvements



Scalability of the bridge service

The bridge service can be easily **scaled horizontally**, with **multiple bridges acting in parallel to support BSC usage**

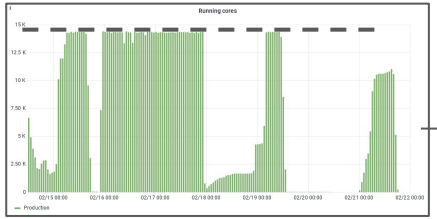
- For these tests, bridge deployed on 2 old to-be-decommissioned WNs at PIC (12 CPU cores and 25 GB RAM each)
- Bridge nodes performance carefully watched during the tests
- Results indicate even this modest setup is sufficient to manage **15k BSC CPU core peaks** with **no saturation detected,**

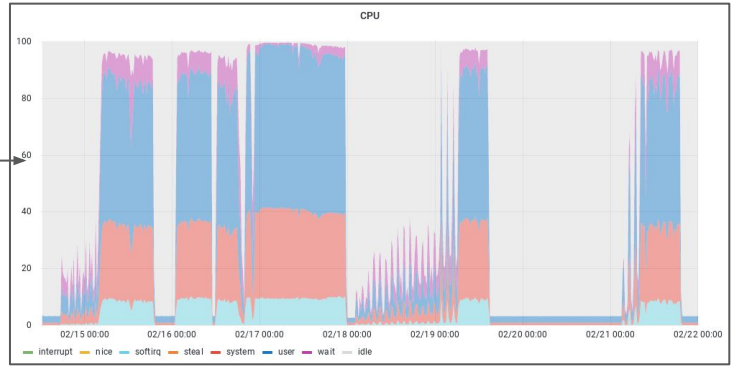


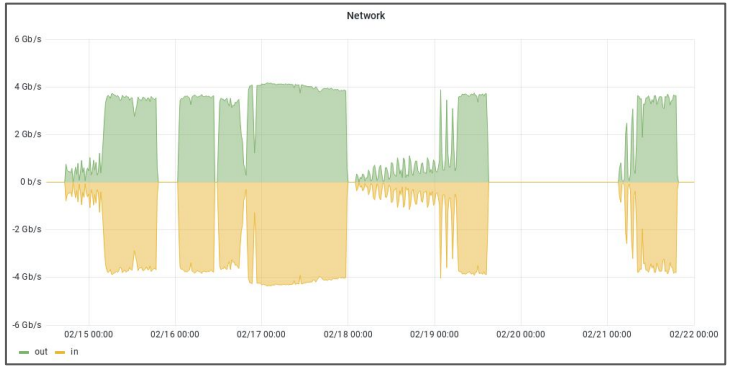
DTS test results (I)

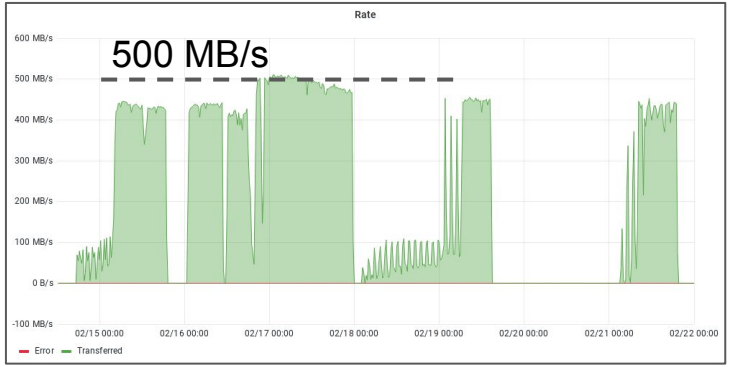
During the tests, the **DTS was deployed on a 8-core 16 GB server with 10 Gbps link**. Running production jobs at maximum scale in BSC (15k CPU cores), the DTS shows some signs of **saturation** (**100% CPU usage**). With **output files at 2 GB**, the effective file transfer to PIC's dCache proceeds at **500 MB/s**.

~15k CPU cores in use





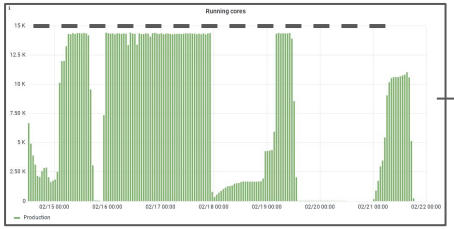


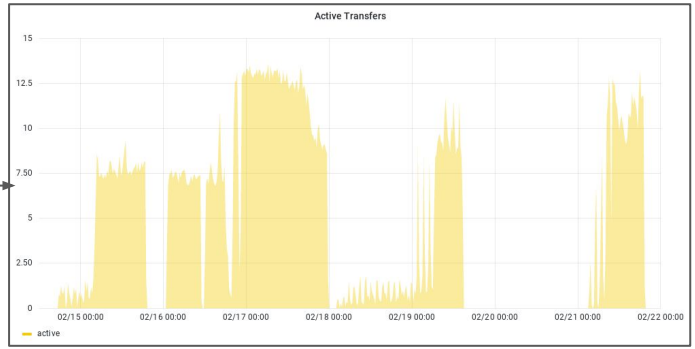


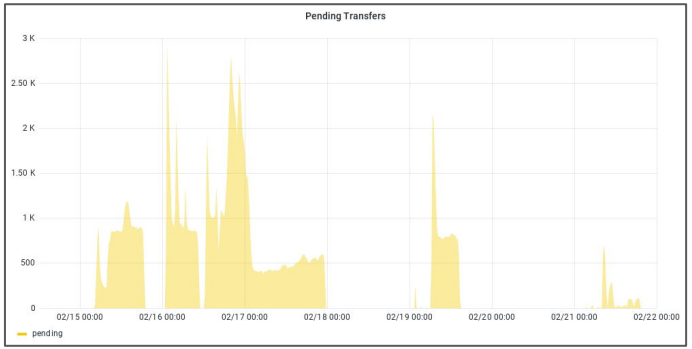
DTS test results (II)

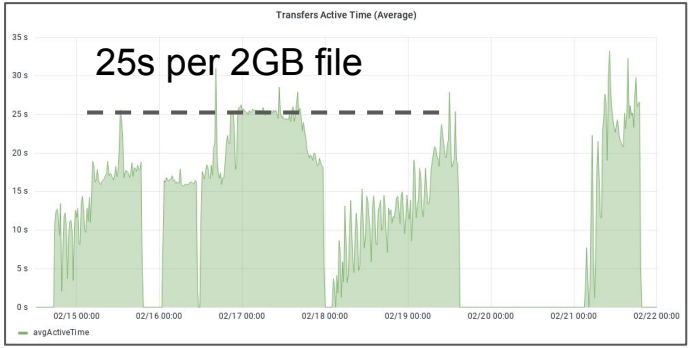
- Producing 2 GB output files, the effective **transfer time** to PIC's dCache is **~25s per job**, while the DTS manages up to **15 concurrent transfers**.
- During each of the test rounds, a large number of jobs start and **end their execution close in time.** This causes an **initial saturation** of the DTS that results in **transfer waiting times up to ~2h** while the backlog is processed. Transfers stay **queued until time out, then jobs are marked as error**.
- As the test round progresses, **job completion times are randomized** and **transfer waiting time is substantially reduced**.

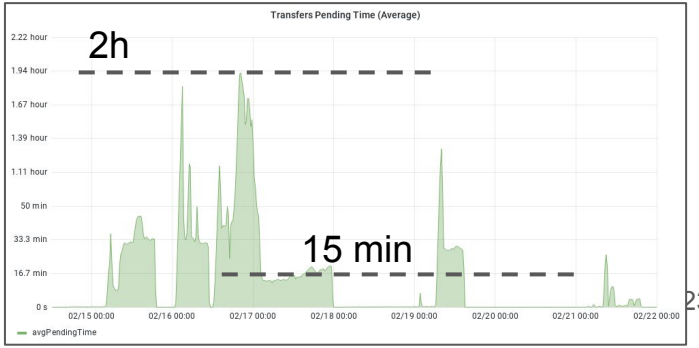
~15k CPU cores in use











- SUMMARY AND NEXT STEPS

PIC
port d'informació
científica

Summary

- Given the network constraints at the HPC, big efforts have been invested in the integration of BSC CPU resources for CMS use
  - HTCondor team: development of split-starter mode
  - CIEMAT team: interface with CMS and BSC, bridge service deployment, configuration, testing, handling of output datasets, etc
  - CMS: Handling of conditions data via files

- A **fully working prototype** has been tested at large scale with realistic CMS workloads

Next steps

Ready to start production work.

Still, further refinements can be applied to our setup to improve automated operations and quality of the service

- Bridge and DTS services to be moved to **more performant HW** at PIC (new DTS server already being procured)
- Expected enhanced **network connectivity** between BSC and PIC for upcoming MN5
  - Spanish NREN to upgrade WAN links to 100 Gbps along 2022
- Improving **monitoring**, **alarms** and implementing **automated recovery** for the services
- **Optimize resource usage efficiency** with closer connection to CMS WMS
  - Moving from vacuum-like glideins to GlideinWMS pilot jobs under consideration
- Possible use of a DTS to manage also input datasets would **increase the usability** of the resources by CMS
  - Full MC production chain, data reprocessing, analysis tasks
- Automated **accounting** algorithms to properly report BSC usage by CMS

- EXTRA SLIDES

PIC
port d'informació
científica

Abstract

Given the growing computing needs of the LHC experiments facing the start of the Run 3 and the High-Luminosity era, it is crucial to gain access to and ensure the efficient exploitation of new resources. The CMS computing ecosystem presents a number of standardized methods for authentication and authorization, access to remotely stored data and software deployment, enabling access to WLCG resources seamlessly. However, incorporating Cloud and HPC resources presents a number of challenges, which generally need to be solved on a case by case basis. The Barcelona Supercomputing Center (BSC), the main Spanish HPC site, represents a particularly difficult case, as severe network restrictions impact the feasibility of many of the aforementioned standardized solutions. This contribution describes a number of actions and novel solutions introduced by the Spanish CMS community in order to facilitate the inclusion of BSC resources into the CMS computing infrastructure for their use by the collaboration. This includes adapting the resource allocation and workload management tools, access to CMS data processing and simulation software, and to remote experimental conditions databases, as well as setting up a transfer service for output data to storage at a nearby WLCG facility. This work summarizes the current status of the integration efforts and also reports on the first experiences with the implemented solutions.

Customizing environment for CMS jobs

- Central **CVMFS** CMS software repository **not accessible from BSC nodes**.
  - Therefore, full repository replication launched from PIC, using cvmfs_preload tool, targeting a CVMFS replica area at BSC shared filesystem (GPFS) mounted with SSHFS for
  - Periodic check for incremental updates

- CMS tasks run at BSC with standard job wrappers, however with a number of environment modifications:
  - Singularity image for CMS job execution container needs to be pre-placed from CVMFS into BSC GPFS
    - Periodically check for changes
  - Container modified to mount BSC's preloaded CVMFS for CMS SW
  - Customized local config files read by CMS tasks for local storage access and job execution (SITECONF)
  - Conditions data read from files, not remote access to DBs, which required modifications at CMS SW level, and also included in the modified SITECONF