HPC use case through PIC - J. Flix

HTCondor Workshop 2024

Amsterdam

23-27 September 2024

PIC port d'informació científica

Exploiting supercomputers for LHC

PIC
port d'informació científica

Data intensive computing with HPC facilities is a **challenge**

- Limited/no network connectivity in compute nodes
- Limited storage for caching input/output event data files - and no edge services (!)
- In practice only run CPU-bound workflows (MC simulation) with little I/O

LHC applications are **not really suited** for HPC

- No large parallelization (no use of fast node interconnects)
- No substantial use of accelerators (GPU) yet

Substantial integration work to make HPC work for HTC

- No one-fit-all solution: each facility is different
- Little effort available in the LHC experiments; in charge of the local communities
- Experiments do not accept pledged HPC resources unless they can be used transparently

No suitable resource allocation model

- We would need a guaranteed share of resources rather than apply for allocations

Then… why?

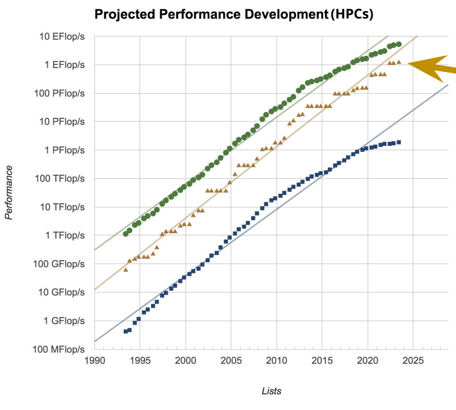
A great **rapidly growing resource** (more than 100x WLCG; WLCG estimated at ~30 PFlops…)

Potential opportunities for **allocations** or **“free” opportunistic** computing usage: can help cover increasing CPU needs despite flat funding

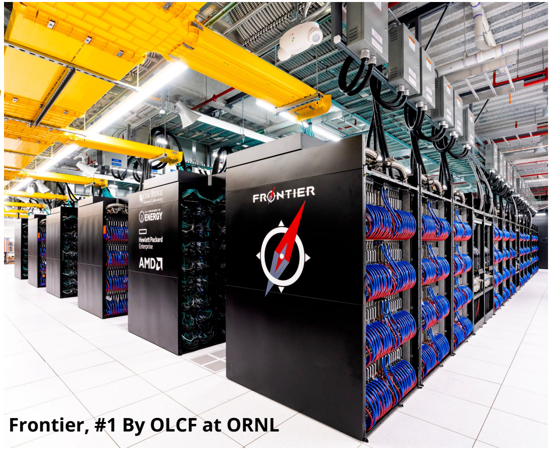
WLCG computing is done economically on the sort of hardware used on the Grid. **National computing priorities** may intend to complement **HTC pledges with HPC resources** at some point soon

Interesting **R&D:** access and use of heterogeneous resources (GPUs, ARMs, POWER)

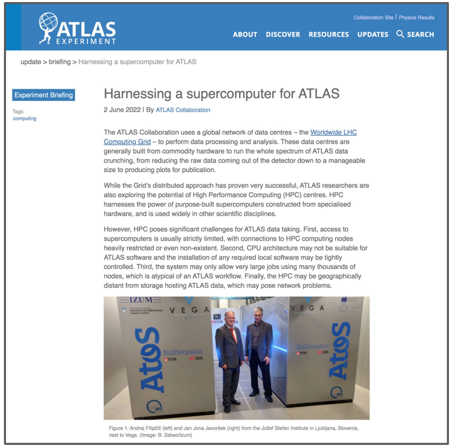
Various barriers for exploitation have been reduced over the years, **some still exist…**



Top 500 list, June 2023



Frontier, #1 By OLCF at ORNL



ATLAS EXPERIMENT

Collaboration Site | Physics Results

ABOUT DISCOVER RESOURCES UPDATES SEARCH

update > briefing > Harnessing a supercomputer for ATLAS

Experiment Briefing

Tags: computing

Harnessing a supercomputer for ATLAS

2 June 2022 | By ATLAS Collaboration

The ATLAS Collaboration uses a global network of data centres – the Worldwide LHC Computing Grid – to perform data processing and analysis. These data centres are generally built from commodity hardware to run the whole spectrum of ATLAS data crunching, from reducing the raw data coming out of the detector down to a manageable size to producing plots for publication.

While the Grid’s distributed approach has proven very successful, ATLAS researchers are also exploring the potential of High Performance Computing (HPC) centres. HPC harnesses the power of purpose-built supercomputers constructed from specialised hardware, and is used widely in other scientific disciplines.

However, HPC poses significant challenges for ATLAS data taking. First, access to supercomputers is usually strictly limited, with connections to HPC computing nodes heavily restricted or even non-existent. Second, CPU architecture may not be suitable for ATLAS software and the installation of any required local software may be tightly controlled. Third, the system may only allow very large jobs using many thousands of nodes, which is atypical of an ATLAS workflow. Finally, the HPC may be geographically distant from storage hosting ATLAS data, which may pose network problems.

Figure 1: Andrej Filipčič (left) and Jan Jona Javoršek (right) from the Jožef Stefan Institute in Ljubljana, Slovenia, next to Vega. (Image: B. Zebec/Izum)

ATLAS briefing on Vega HPC, June 2022

Integration of BSC CPU resources

PIC port d'informació científica

In 2020 BSC designated LHC computing as a **strategic project**

- Agreement promoted by WLCG-ES community and funding agency

Allocations* can use **'reserved' resources** to strategic projects

- Currently: ~90M coreHours/year

* Submission of proposals for CPU time allocation every 4 months

Potentially, very **significant contribution** for LHC computing in Spain

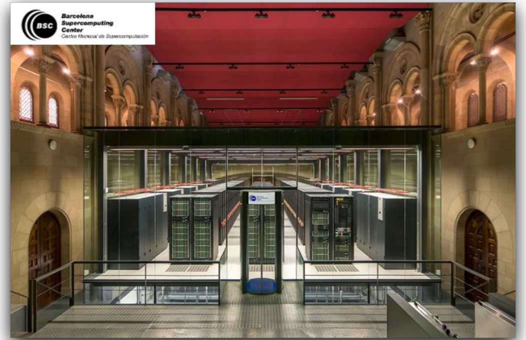
- Comparable e.g to all ATLAS+CMS+LHCb <u>simulation needs</u> in the country

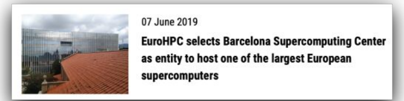
Integration of BSC CPU resources

BSC - Barcelona Supercomputing Center

- Largest HPC center in Spain
- **MareNostrum 4** (*MN4*) general-purpose cluster:
  - 11.5 Petaflops (166k CPU cores), 390 TB RAM, 24 PB disk local SSD disk
  - SLURM as batch system, SUSE Linux Enterprise as OS
  - 15 PB GPFS as storage back-end (mounted on login/compute nodes)
- **MareNostrum 5** (*MN5*: ~17xMN4, ~200 petaflops)
  - One of Europe's first pre-exascale supercomputers
  - 730k CPU cores - 112 cores/node - 250 PB GPFS disk storage

https://eurohpc-ju.europa.eu/about/our-supercomputers_en#marenostrum-5





BSC imposes very **restrictive network connectivity** conditions

- No incoming *or outgoing* connectivity from compute nodes
- Only incoming SSH/SSHFS communication through login nodes
- A shared disk (GPFS) mounted on compute nodes and login machines - accessible from outside via sshfs
- No services can be deployed on edge/privileged nodes

Use of the BSC resources

PIC port d'informació científica

Services installed at the Spanish WLCG sites to access and exploit BSC resources:

- **PIC Tier-1**: **2x ARC-CEs** for both **ATLAS** and **LHCb**, and **custom-made gateway** for **CMS**
- **IFIC Tier-2**: **1x ARC-CE** for **ATLAS**
- **UAM Tier-2**: **1x ARC-CE** for **ATLAS**

Submission of **proposals for time allocation** every 4 months

- 3x proposals for **ATLAS** (A. Pacheco-IFAE, S. González-IFIC, J. del Peso-UAM)
- 1x proposal for **CMS** (J. Flix-CIEMAT)
- 1x proposal for **LHCb** (X. Vilasis-Ramon Llull)

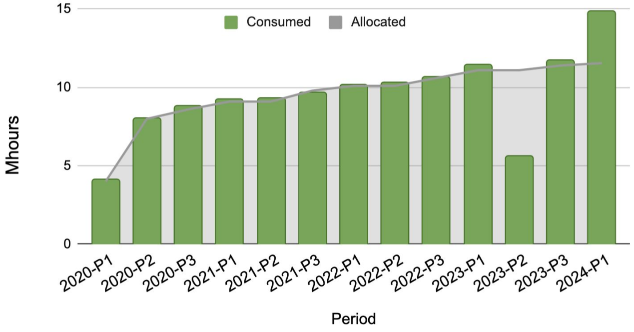
Since 2020, **~75 allocation proposals have been submitted** and approved by BSC

- Allocations sum up **~202 million CPU hours**
- We have utilized ~**201.5 million CPU hours**

Taking into account the period length and CPU power at BSC, this utilization corresponds to an average installed capacity of approximately 105 kHS06, representing around **48% of the average Grid resources deployed in Spain** for the LHC experiments in the period 2020-2024

Use of the BSC resources

ATLAS

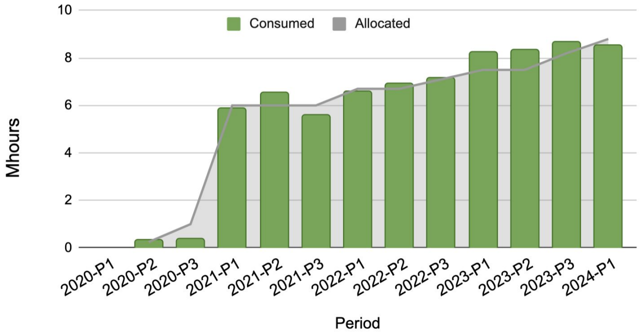


[PIC ARC-CE gateway]

[IFIC ARC-CE gateway]

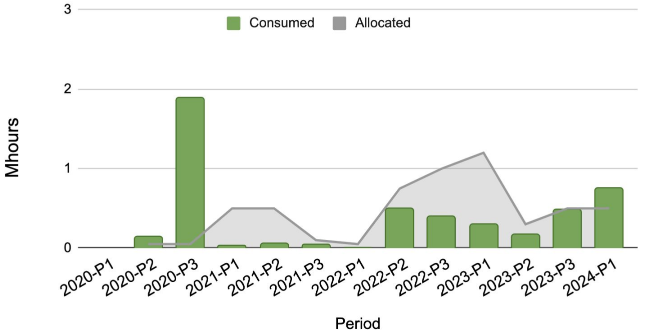
[UAM ARC-CE gateway]

CMS

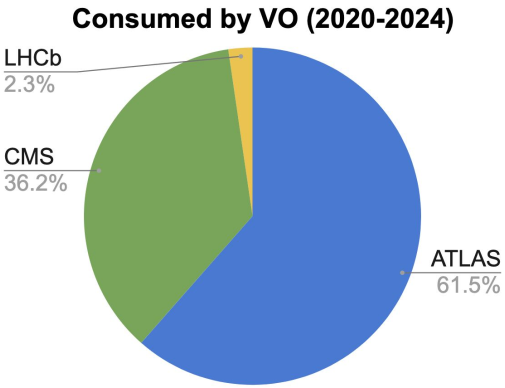


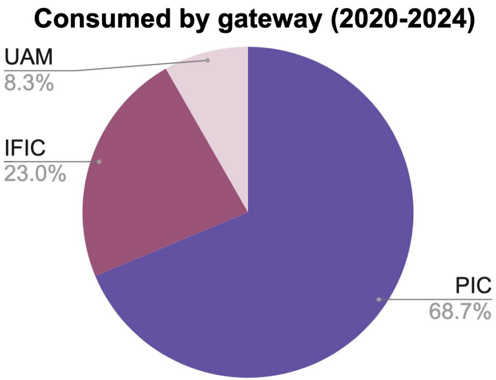
[PIC custom-made gateway]

LHCb

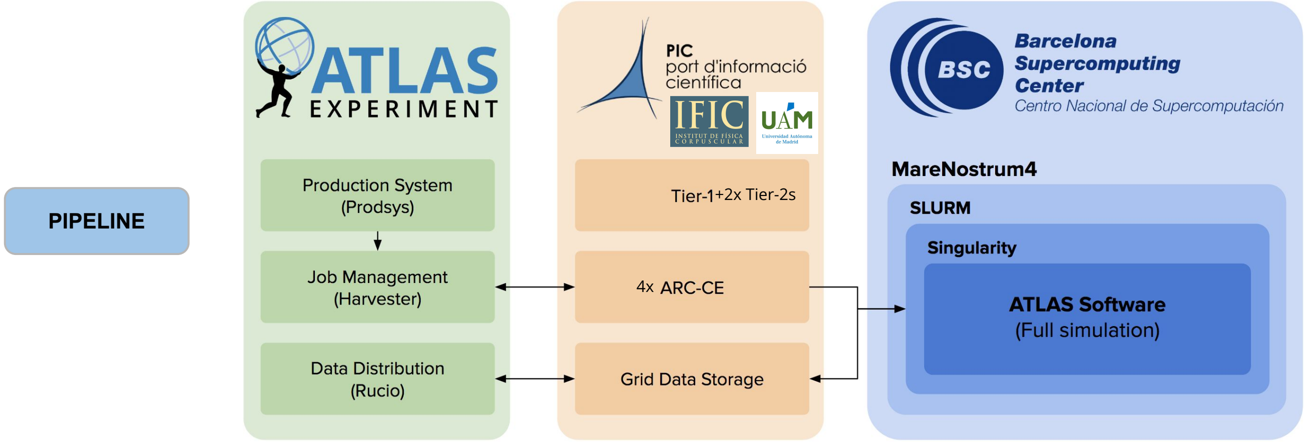


[PIC ARC-CE gateway]





Use of the BSC by ATLAS



Submitting **ATLAS** payloads to BSC since 2018, in production since 2019

Using four **ARC-CEs** in Spain to interconnect MareNostrum and ATLAS production system

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

~33 million CPU hours used at BSC **last year** by ATLAS through these gateways

→ **At CHEP2021 proceedings (link)**
→ **At CHEP2023 (link)**

Use of the BSC by LHCb

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

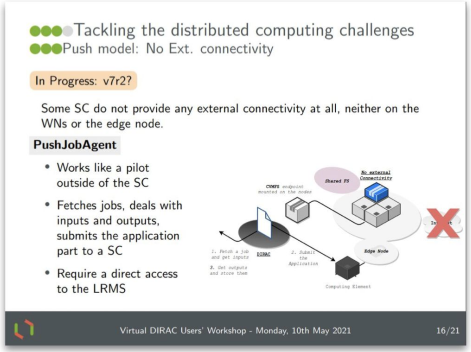
~1-5 million CPU hours used at BSC **last year** by LHCb through this gateway (simulations)

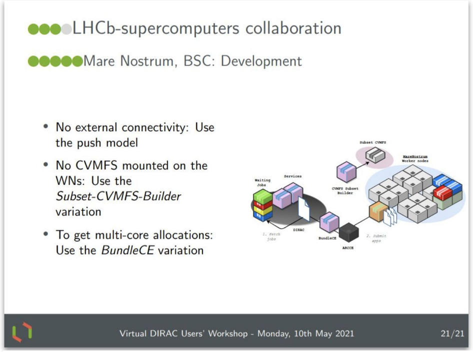
LHCb

Node ES-PIC — SUM Wallclock Work (cores * HS23 hours) by Submit Host and Month (Custom VOs)

Submit Host	Dec 2023	Jan 2024	Feb 2024	Mar 2024	Apr 2024	May 2024	Jun 2024	Total	Percent
ce13.pic.es:9619/ce13.pic.es-condor	8,821,362	9,589,883	10,378,281	9,093,341	7,357,221	10,547,864	2,846,309	58,634,261	44.82%
ce14.pic.es:9619/ce14.pic.es-condor	6,233,465	7,101,988	10,398,485	9,270,233	7,298,901	10,344,778	2,774,184	53,422,035	40.84%
gsiftp://arc-ce02.pic.es:2811/jobs	0	0	0	0	0	0	0	0	0%
https://arc-ce02.pic.es:8443/arex	1,758,237	2,198,497	3,377,220	2,800,666	2,308,863	4,730,258	1,592,607	18,766,348	14.34%
Total	16,813,064	18,890,369	24,153,986	21,164,240	16,964,985	25,622,900	7,213,101	130,822,644	
Percent	12.85%	14.44%	18.46%	16.18%	12.97%	19.59%	5.51%		

1 - 4 of 4 results ‹ 1 › Number of rows per page 30





+ **DIRAC X** developments

A.Boyer, "Integrating DIRAC workflows in Supercomputers" [link]

Use of the BSC by CMS

Reminder: BSC imposes very **restrictive network connectivity** conditions

- No incoming *or outgoing* connectivity from compute nodes
- Only incoming SSH/SSHFS communication through login nodes (*ui*)
- A shared disk (GPFS) mounted on compute nodes and login machines - accessible from outside via sshfs
- No services can be deployed on edge/privileged nodes

This ***was*** a **major obstacle for CMS** workloads:

- Pilot with late binding model execution of payloads
  - Workload management system (glideinWMS - HTCondor services)
- Access to external services
  - Application software (CVMFS) & Conditions data (FrontierDB)
- Consuming and producing experiment data
  - Input/output data files (Storage Elements)

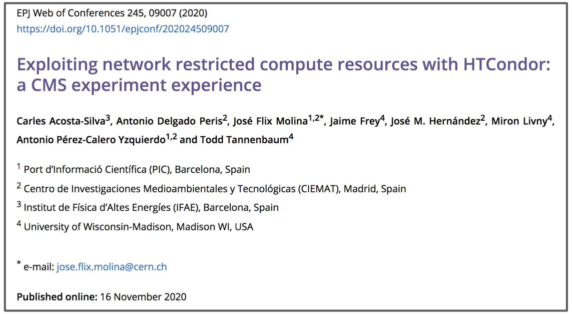
Use of the BSC by CMS: solutions

HTCondor 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] and [CHEP21]
- Requires a **bridge service at PIC** to connect CMS WMS and BSC

A collaboration was formalized during the September’18 RAL HTCondor workshop

[Miron Livny, Todd Tannenbaum, Jaime Frey, Antonio Pérez-Calero, Carles Acosta, José Flix]



EPJ Web of Conferences 245, 09007 (2020)
https://doi.org/10.1051/epjconf/202024509007

Exploiting network restricted compute resources with HTCondor: a CMS experiment experience

Carles Acosta-Silva[3], Antonio Delgado Peris[2], José Flix Molina[1,2*], Jaime Frey[4], José M. Hernández[2], Miron Livny[4], Antonio Pérez-Calero Yzquierdo[1,2] and Todd Tannenbaum[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] University of Wisconsin-Madison, Madison WI, USA

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

Published online: 16 November 2020

Use of the BSC by CMS: solutions

CMS software (**CMSSW**) deployed to BSC environment via **CVMFS pre-loaded replica**

Conditions data accessed via double reverse **ssh tunnels**

Developed and then **operate** a **custom data transfer service** (DTS) for output data migration from BSC to PIC storage

Setup local environment configuration for CMS tasks (e.g. where to write output data)

BSC currently running **MC GEN jobs in TaskChain** mode → **pile-up further added in PIC** to produce final MC samples

Solutions working at scale presented at **CHEP23**

Use of the BSC by CMS: solutions

CMS software (**CMSSW**) deployed t[…]
replica

Conditions data accessed via doub[…]

Developed and then **operate** a **cu**[…] […]ut
data migration from BSC to PIC sto[…]

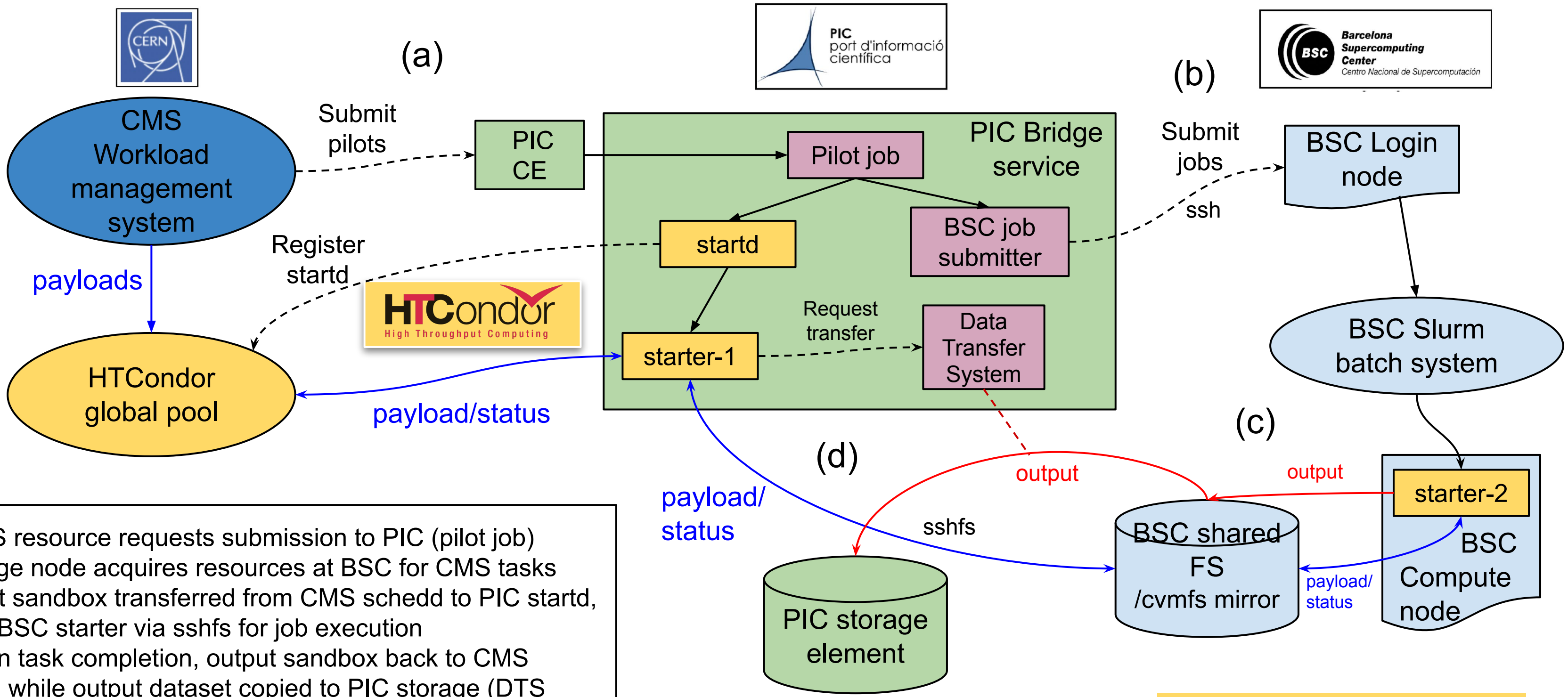
Setup local environment configur[…] […]tput
data)

Two modes of production: **TaskChain** mode, where different steps are performed asynchronously, writing the intermediate data to the global CMS storage (each TaskChain step contains one or more steps of the simulation, GEN, SIM, DIGI, PUMIX, RECO), and **StepChain**, where all the simulation steps are executed in a single job.

BSC currently running **MC GEN jobs in TaskChain** mode → **pile-up further added in PIC** to produce final MC samples

Solutions working at scale presented at **CHEP23**

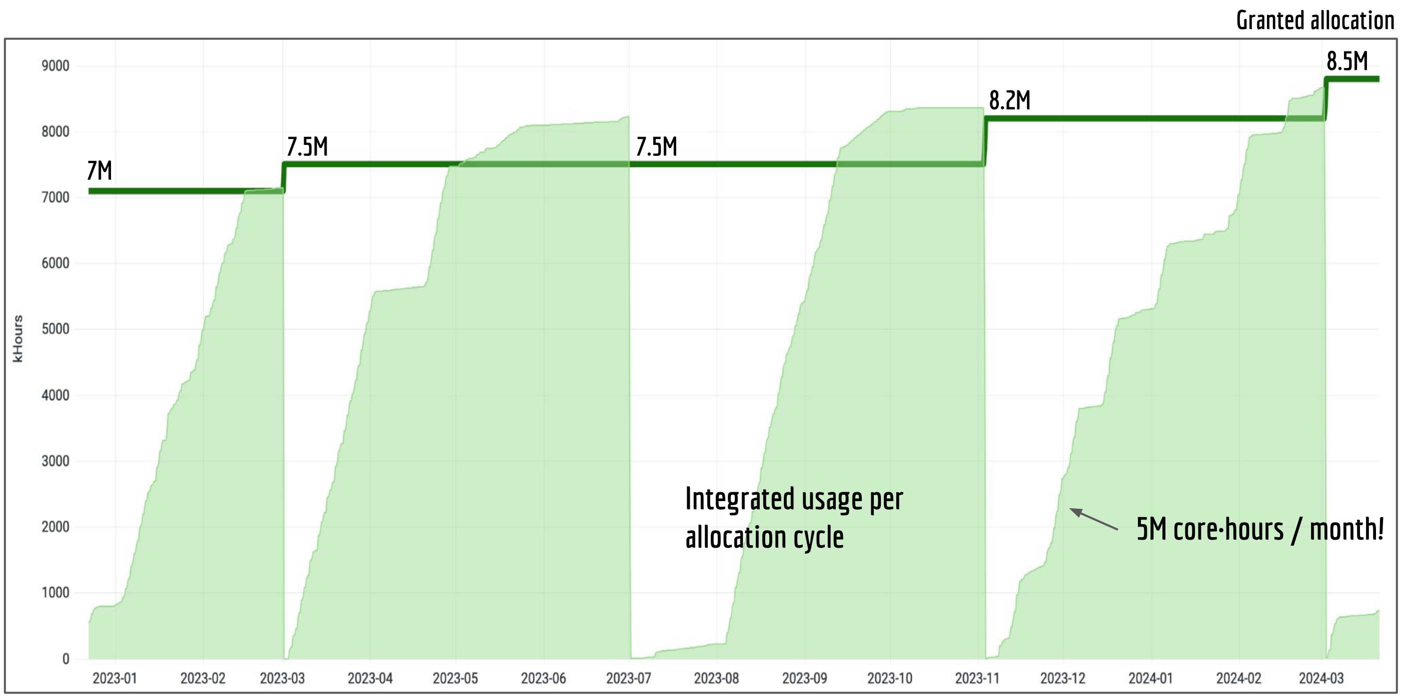
Use of the BSC by CMS: the solution (!)



(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**

Use of the BSC by CMS: scale



*** Peaks of 12.5k cores used by CMS @ BSC (~8% of BSC MN4 CPUs)**

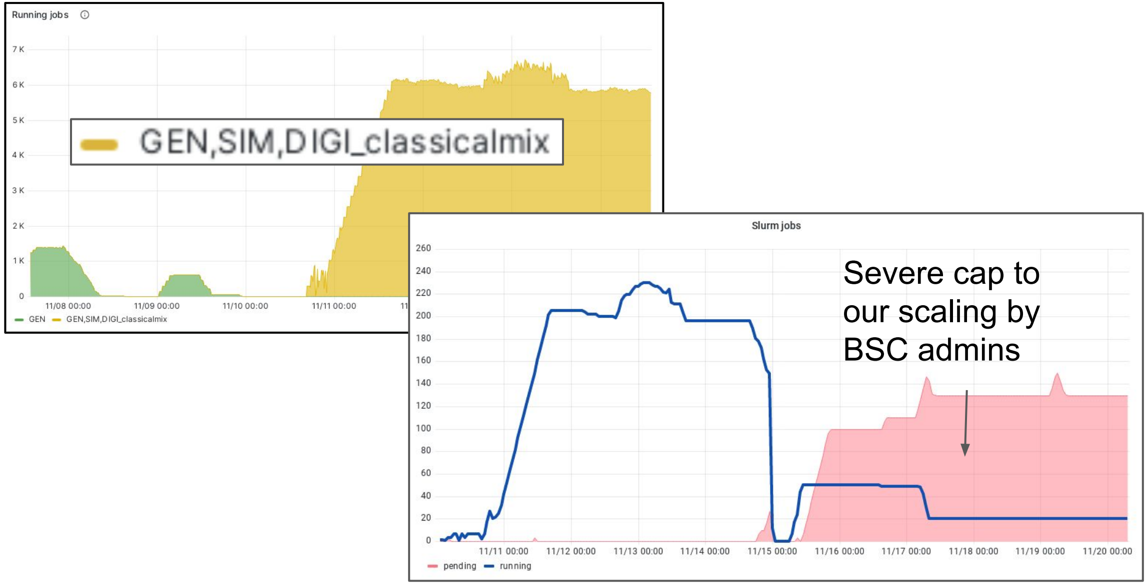
→ At CHEP2021 proceedings (link)
→ At ISGC 2022 (link)
→ At HTCondor WS 2022 (link)
→ At CHEP2023 (link)

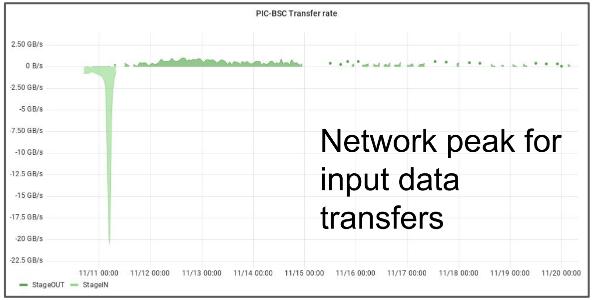
Use of the BSC by CMS: Special workflows

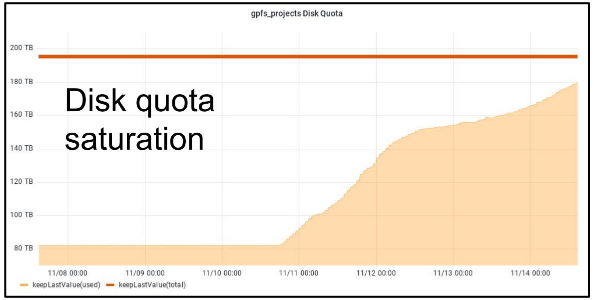
Tried **pre-mixed samples production** at BSC (MN4): using the individual pileup event samples to produce a sample (called pre-mix) where events are built, and the number of PUC collisions follows the data profile.

Input data was successfully copied from PIC to BSC (first use of DTS for input files at scale), importing data really limited by BSC bandwidth saturation (10 Gbps).

Consequences: rapidly growing utilization of the local disk quota. **IO-intensive jobs saturating GPFS**, got complaints from BSC admins, and our **max node quota was severely reduced… but it worked for a while!**







New MN5 facility

New MN5 facility

Recently commissioned (OS changes, apptainers, ...). In general, we are seeing the **same level of difficulties to operate in MN5**

Increase of resources at BSC, but still some **limitations**: e.g. maximum disk space for allocation of 500 TB (impossible to fit 1 PB of pile-up CMS sample)

Managing also input datasets by all of the experiments would increase the usability of the MN5 resources: i.e. data re-processing and analysis tasks

Enhanced **network connectivity**: BSC WAN increased (10 Gbps → 200 Gbps)

- In general, better tools to transfer data between HPCs and HTCs is needed

Potential for **GPU resources** exploitation through MN5 facility

In order to reduce Grid project costs, the **CPU usage of BSC is part of the Spanish WLCG pledge from 2024 on** (even if we are not yet running all type of workflows)

Using pile-up samples at BSC

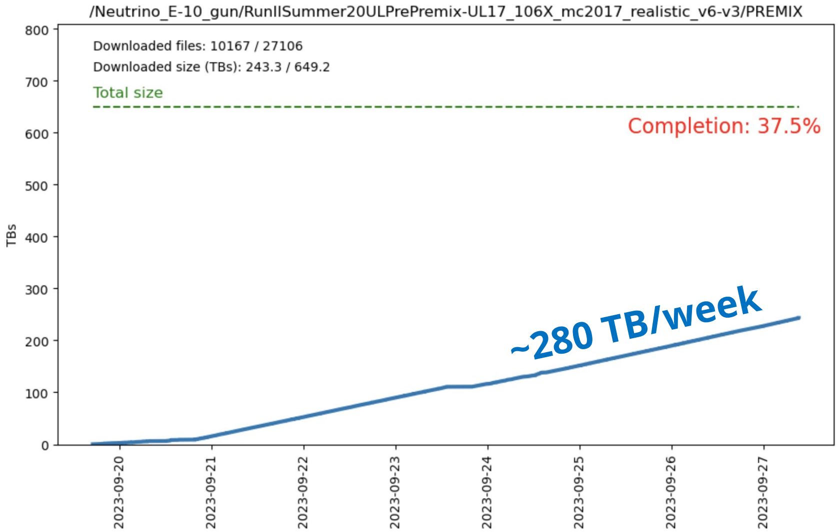
PIC port d'informació científica

In the last allocation, we asked for **2PB disk space in MN5** and… **it was approved!**

This would allow us to **run the complete MC workflow chain at BSC**, reading pre-placed pile-up samples in BSC disks

We are currently **transferring pile-up samples** used for legacy MC campaigns

Load balanced xrdcp's from CERN|FNAL to BSC transfer nodes (4), using sshfs mounted areas at PIC server



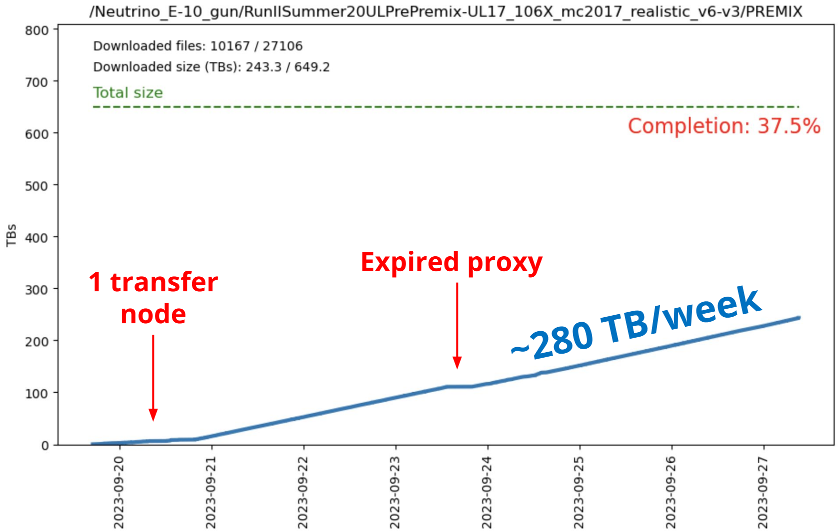
Using pile-up samples at BSC

In the last allocation, we asked for **2PB disk space in MN5** and… **it was approved!**

This would allow us to **run the complete MC workflow chain at BSC**, reading pre-placed pile-up samples in BSC disks

We are currently **transferring pile-up samples** used for legacy MC campaigns

Load balanced xrdcp's from CERN|FNAL to BSC transfer nodes (4), using sshfs mounted areas at PIC server



Conclusions

PIC port d'informació científica

A lot of **great work has been done** to exploit HPCs in WLCG. In particular, a difficult context has been addressed to use BSC CPU resources by ATLAS, CMS and LHCb

So far the exploitation since 2020 represents around 48% of the average grid resources deployed in Spain - **compatible with the 50% target** for exploitation set for 2024 on

The work done to integrate this resource took **many FTE efforts** from the WLCG Spanish community and from international teams (HTCondor and experiment frameworks)

Current challenges on the software side to **exploit different architectures** (GPU). Door for opportunities to test/use these type of resources at MN5

Some HPC facilities look like they are **becoming more friendly to HEP**, at least in terms of accessibility. However, we are seeing the **same level of difficulties in MN5**

Since HPC facilities are constantly being designed... Can WLCG present a **united front** and **have a voice** to influence?

- Joint ECFA-NuPECC-APPEC (JENA) workshop resulting in a Working Group to focus on this, including Particle Physics, Nuclear Physics, and Astroparticle Physics communities (data intensive sciences)

Bedankt!

PIC
port d'informació
científica