Integrating BSC Computing Infrastructure into ATLAS/CMS/LHCb

J. Flix

CPAN Ingenio 2010

PIC port d'informació científica

XVI CPAN Days, Madrid
19-21 November 2024

IFAE Ciemat RES Red Española de Supercomputación España digital 2026 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

The context: HPC integration in LHC

PIC
port d'informació científica

Lots of **supercomputing resources** (HPC) available and more on the way

- Increasing use of HPC resources in WLCG
- Pressure from funding agencies to use those resources

LHC applications are not really suited for HPC

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

Challenging integration in LHC distributed computing

- No one-fit-all solution: <u>each HPC facility is different</u>
- HPCs **well suited for CPU-intensive workflows** (e.g. MC simulation production)
- Challenging for **data-intensive workflows**
  - Limited/no network connectivity in compute nodes
  - Limited storage for caching input/output event data files

WLCG does not accept **pledged HPC resources** unless they can be used transparently

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

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

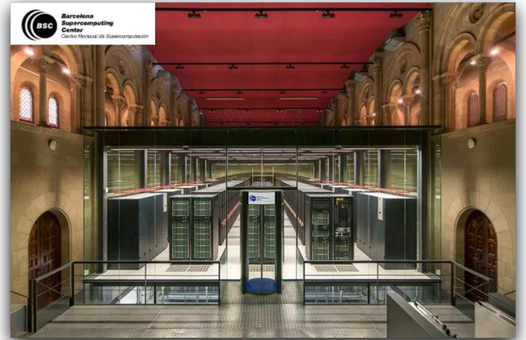
- Comparable e.g to all ATLAS+CMS+LHCb <u>simulation needs</u> in the country

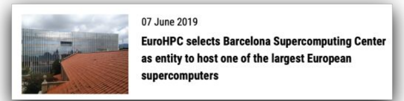
Integration of BSC CPU resources

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
  - 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

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

- **PIC Tier-1 / IFAE Tier-2**: **3x ARC-CEs** for both **ATLAS [2x]** 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)

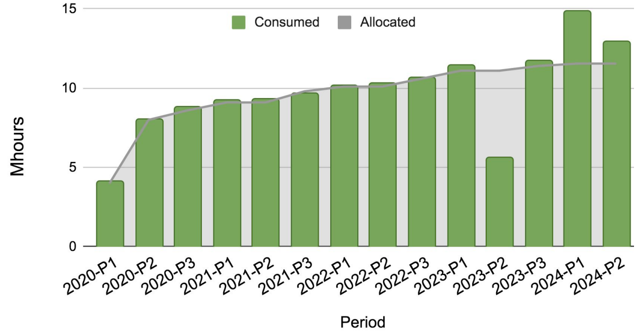
Since 2020, **~80 allocation proposals have been submitted** and approved by BSC

- Allocations sum up **~223.2 million CPU hours**
- We have utilized ~**226.3 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

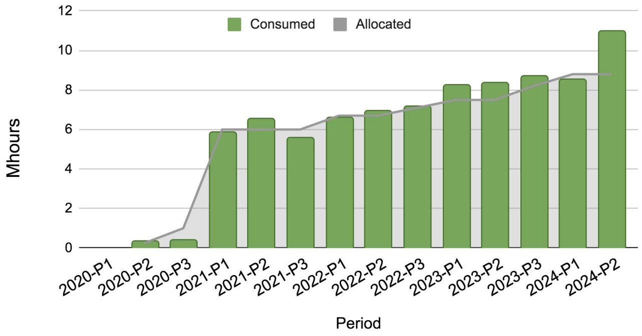


[PIC ARC-CE gateway]

[IFIC ARC-CE gateway]

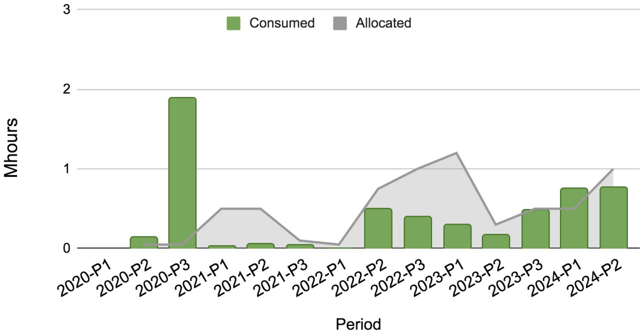
[UAM ARC-CE gateway]

CMS

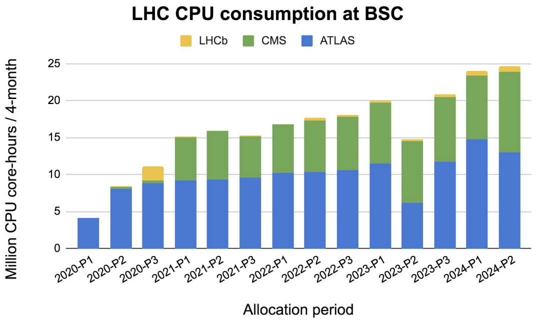


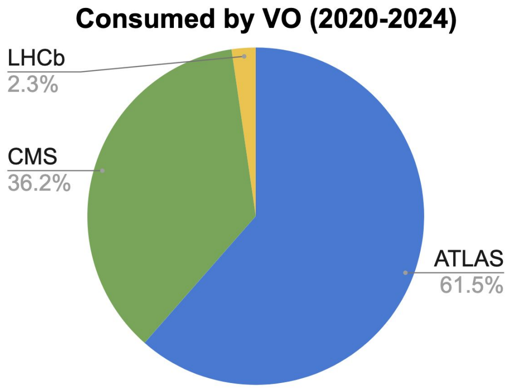
[PIC custom-made gateway]

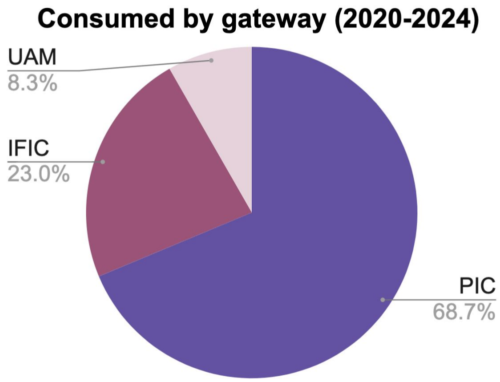
LHCb



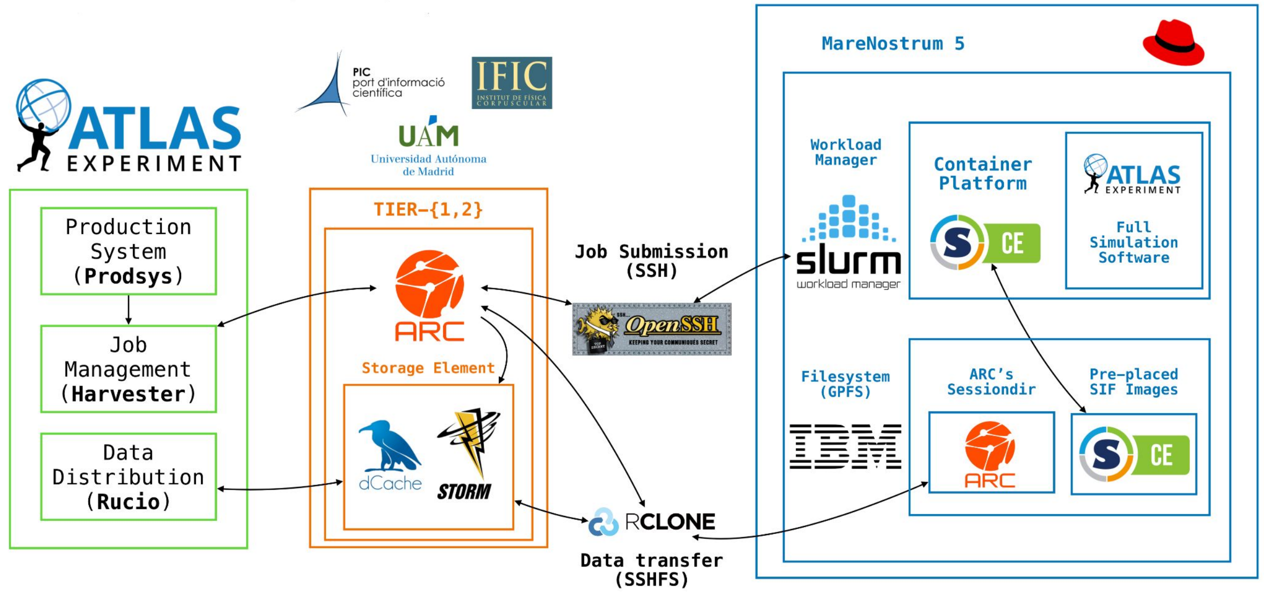
[PIC ARC-CE gateway]







Use of the BSC by ATLAS



Submitting **ATLAS** payloads to BSC since 2018, <u>in production since 2019</u>

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

<u>Only simulation workflow</u> validated - singularity containers, pre-placed at MareNostrum GPFS

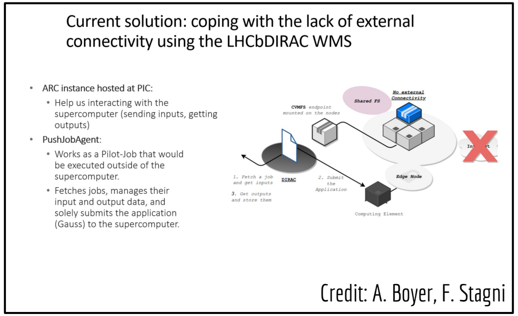
~33 million CPU hours used at BSC **last year** by ATLAS through these gateways

→ **At: CHEP2021 (link) | CHEP2023 (link) | CHEP2024 (link)**

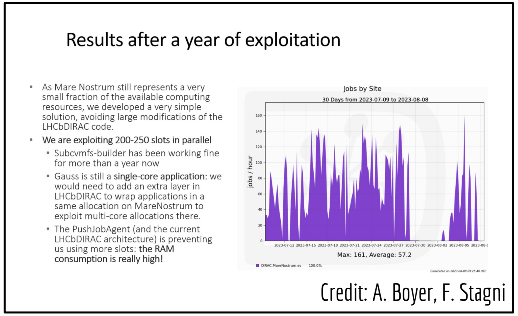
Use of the BSC by LHCb

LHCb uses similar technical implementations as ATLAS (**ARC-CE05.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)



Credit: A. Boyer, F. Stagni



Credit: A. Boyer, F. Stagni

- 2 *PushJobAgents* submit parts of jobs in MN5 (**~350 single-core jobs**)
- regular uploads of large Singularity containers with content extracted from CVMFS… the lack of CVMFS remains a significant constraint
- The scenario is far from ideal, as it consumes a significant amount of memory on the voboxes @CERN where they are installed
- The BSC CPU **usage is at ~15% levels according to planification**

Solutions:

- **Gaussino**: a multi-threaded MC Simulation application, intended to replace Gauss, the current MC Simulation application next year(s)
- **New *PushJobAgent*** capable to submit complete jobs, which will be part of DIRAC v9 (Q2-2025) - though this won't solve utilization…

→ **DIRAC** documentation

Use of the BSC by CMS

PIC port d'informació científica

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

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)

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

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.

Use of the BSC by CMS: solutions

PIC port d'informació científica

CMS software (**CMSSW**) deployed to
replica

Conditions data accessed via doub

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

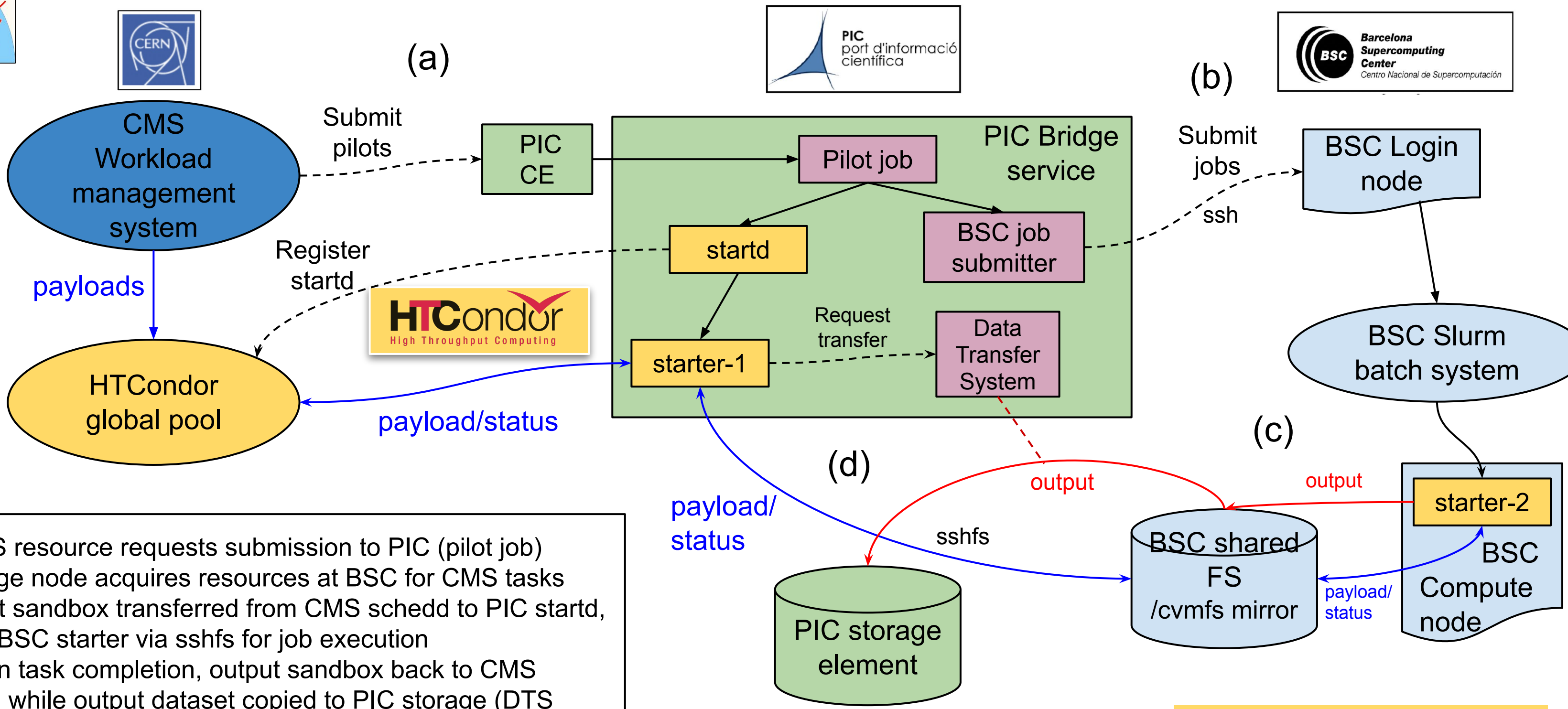
Setup local environment configur tput
data)

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]** **and** **[CHEP24]**

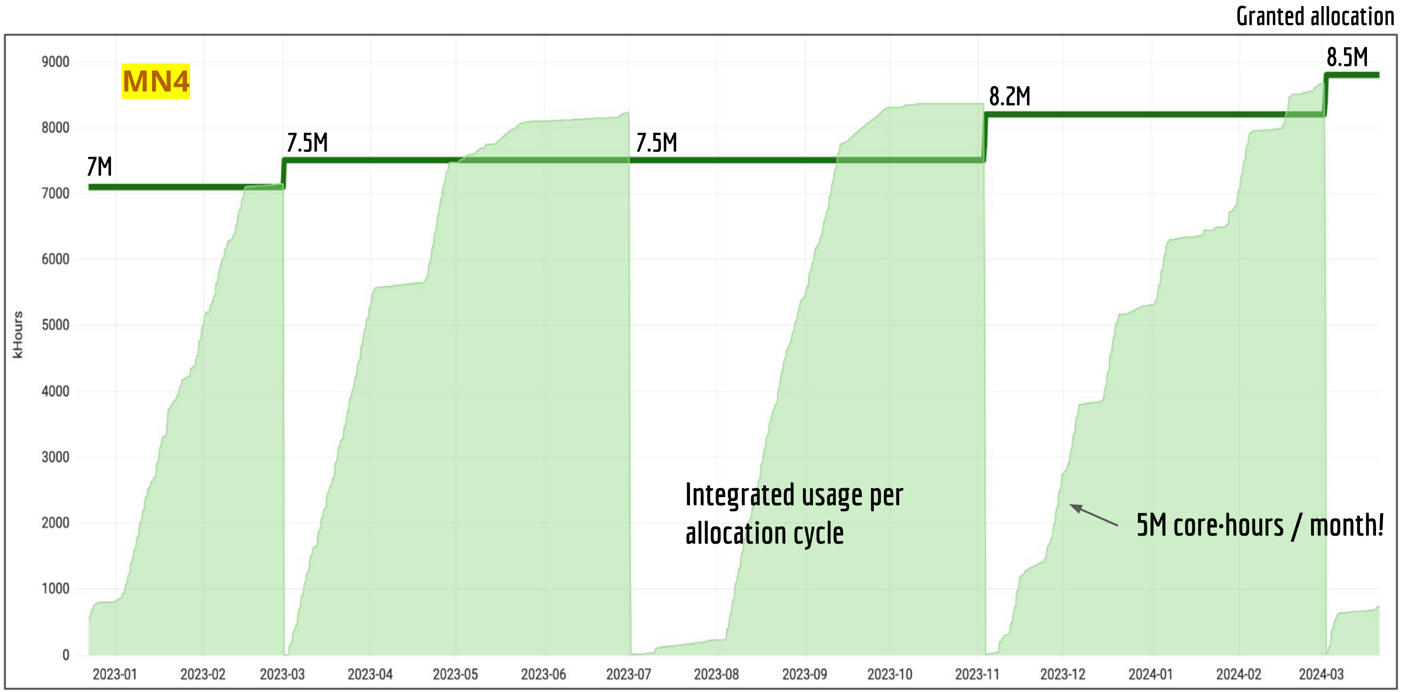
Use of the BSC by CMS: the solution (!)



(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)**

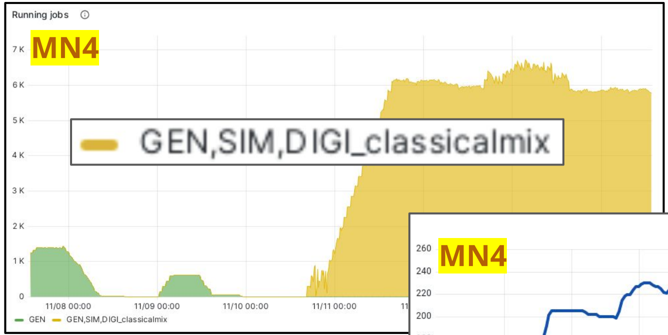
→ **At CHEP2021 proceedings (link)**
→ **At ISGC 2022 (link)**
→ **At HTCondor WS 2022 (link)**
→ **At CHEP2023 (link)**
→ **At CHEP2024 (link)**

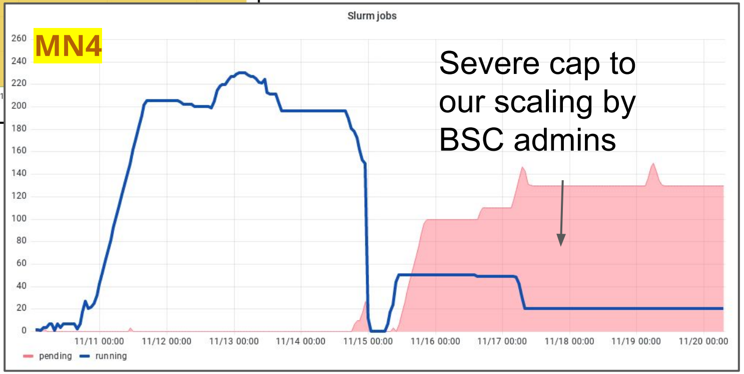
Use of the BSC by CMS: Special workflows

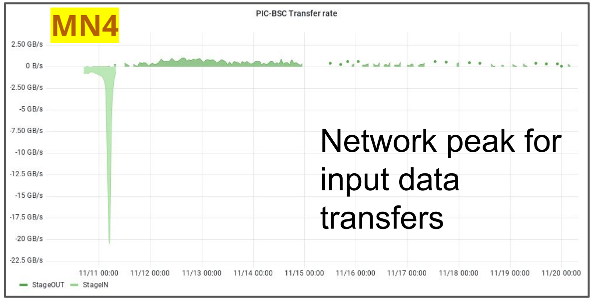
Tried **pre-mixed samples production** at BSC (MN4): using the individual pileup event samples to produce a sample (called pre-mix) where events are built, and the number of PU collisions follows the data profile

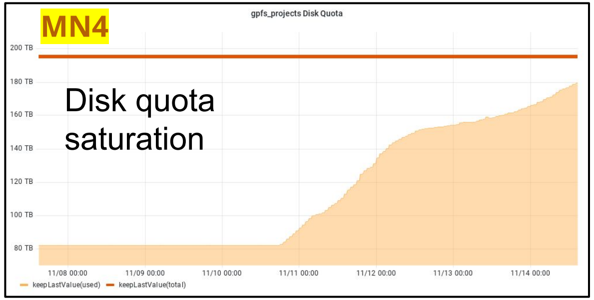
Input data was successfully copied from PIC to BSC (first use of DTS for input files at scale), importing data really limited by BSC bandwidth saturation (10 Gbps)

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

PIC
port d'informació
científica

New MN5 facility

PIC
port d'informació
científica

MareNostrum 5 (MN5: ~17xMN4, ~200 petaflops)

- One of Europe's first pre-exascale supercomputers (#8 in Top500 list June 2024)
- General purpose partition: 730k CPU cores - 112 cores/node - 250 PB GPFS disk storage
- ~25 HS23/core ➟18M HS23 (about the whole WLCG CPU capacity!)
- Commissioned ~June 2024 → in operation since September 2024

Same limitations as in MN4

- No outbound network connectivity from compute nodes 😖

New opportunities

- Large Accelerated Partition: 1120 nodes: 64 CPU cores, 4 x Nvidia Hopper GPUs
- 250 PB GPFS disk storage
- Increased WAN bandwidth (400 Gbps)

We **transitioned** to use MN5 in a few weeks for ATLAS, CMS and LHCb in period 2024-P1

MN5 available resources

PIC port d'informació científica

Increased x5 CPU capacity for MN5 from MN4, **how much can we effectively use?**

Current availability of the MN5 cluster indicates **we can allocate resources as needed**

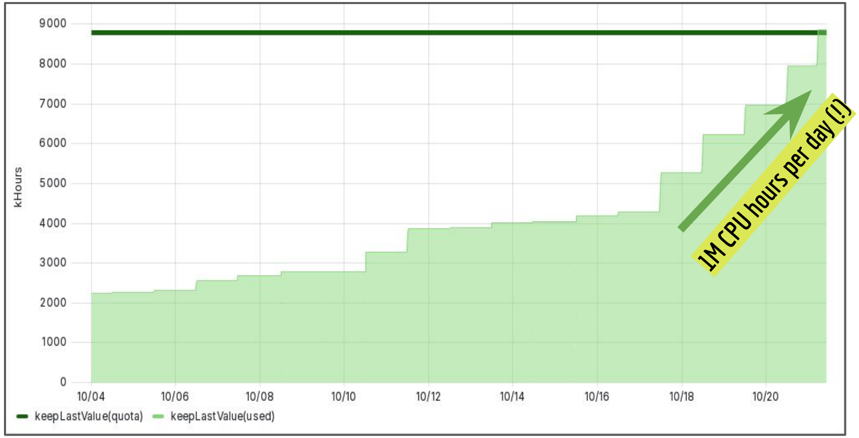
Example: MN5 use for CMS

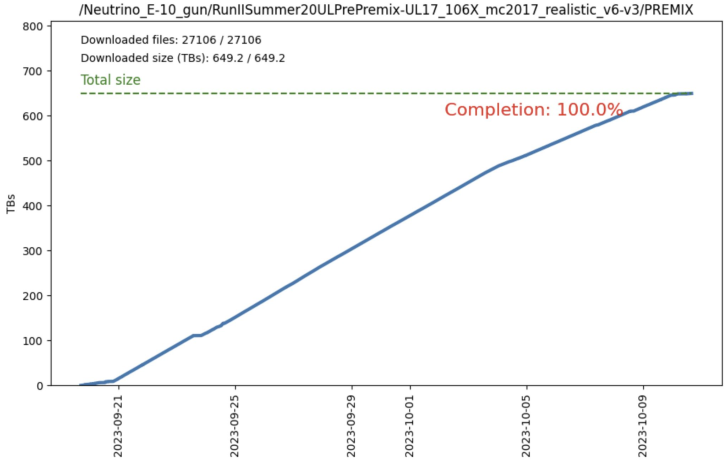
Max. allowed use of 360 nodes → **46k CPU cores** (~1.2 MHS23!) → MC sim. ~15 Gbps BSC→PIC

~6.5% of MN5

Understanding of our setup **scalability limitations**, along with selection of most suitable CMS workloads (MinBias generation at 16 CPU cores per job) → **1M CPU hours per day (!)**

2 PB space allocation in MN5: **copied two "big" pile-up event datasets (~650TB/each) to BSC** to run MC simulation with all steps including pile-up mixing ← StepChain mode





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

- Mostly MC simulations. Supporting other workflows with input data foreseen

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, WLCG is pushing for a **united front** to **have a voice** to influence future HPC designs:

- 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)

Acknowledgements

The authors of this work express their gratitude to the PIC, CIEMAT, IFAE, IFIC, UAM, URL, IFCA and BSC teams for their support in the integration of BSC resources for LHC activities in the Spanish region. This work is partially financed by the Spanish Ministry of Science, Innovation and Universities MCIN AEI/10.13039/501100011033 through the grants with references PID2022-142604OB-C21/C22, PID2022-136323OB-C21/C22, and PID2022-139519OB-C21, which include FEDER funds from the European Union, and also the Catalan government under contract 2021 SGR 00574.

Thanks!

Questions?

Authors: J. Flix, A. Pacheco, C. Acosta, A. Pérez-Calero, G. Merino, V. Acín, J. Casals, S. González, J. Salt, F. Sánchez, J. M. Hernández, A. Delgado, J. del Peso, P. Collado, I. Cabrillo, F. Matorras, J. Saborido, A. Fernández, X. Vilasís, F. Martinez, J. Cuevas, M. J. García