Exploitation of the MareNostrum 4 HPC using ARC-CE

Carlos Acosta-Silva, José del Peso, Esteban Fullana, Santiago González de la Hoz, Andrés Pacheco Pages, José Salt, Javier Sánchez

vCHEP2021 - Wednesday 19 May 2021

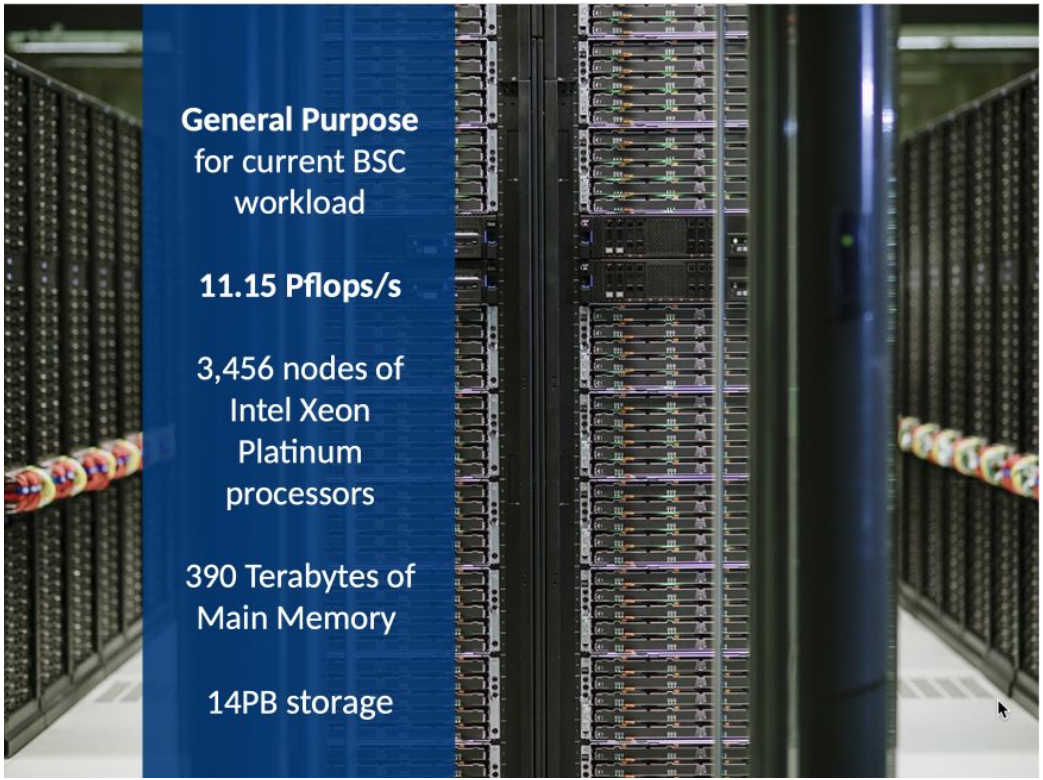
Introduction

- Spain is contributing to the WLCG grid since the first years of the LHC's commissioning.
- We have a Tier-1 and several federated Tier-2 for ATLAS, CMS, and LHCb.
- Pledges, availability, and reliability have been accomplished.
- Now, we are entering another economic cycle and much concern has been raised about the funding continuity of computing resources for the LHC.
- The primary motivation for integrating the HPC centers in Spain LHC computing is to reduce the cost and take advantage of the new massive computing infrastructures.

The HPC situation in Spain

- In Spain, there is a distributed network of supercomputing centers spread throughout the country called RES.
- To contribute to the computation of the LHC, we have used the centers located in Cáceres, Madrid, and Barcelona.
- But by far, the most powerful machine is MareNostrum 4, located in the Barcelona Supercomputing Center (BSC).
- The LHC Computing has been approved as strategic project in the BSC.
- MareNostrum 4 is planning an upgrade in 2022 with EuroHPC pre Exascale funding. The new machine will be MareNostrum 5.

MareNostrum4 Picture



- Each node has two Intel Xeon Platinum chips, each with 24 processors, amounting to a **total of 165,888 processors** and a main memory of **2 GB RAM per processor**.
- Batch system: **SLURM**
- Operating system: **SUSE Linux Enterprise Server 12 SP2**
- Shared file system: **GPFS**

Integration of Spanish HPC Resources

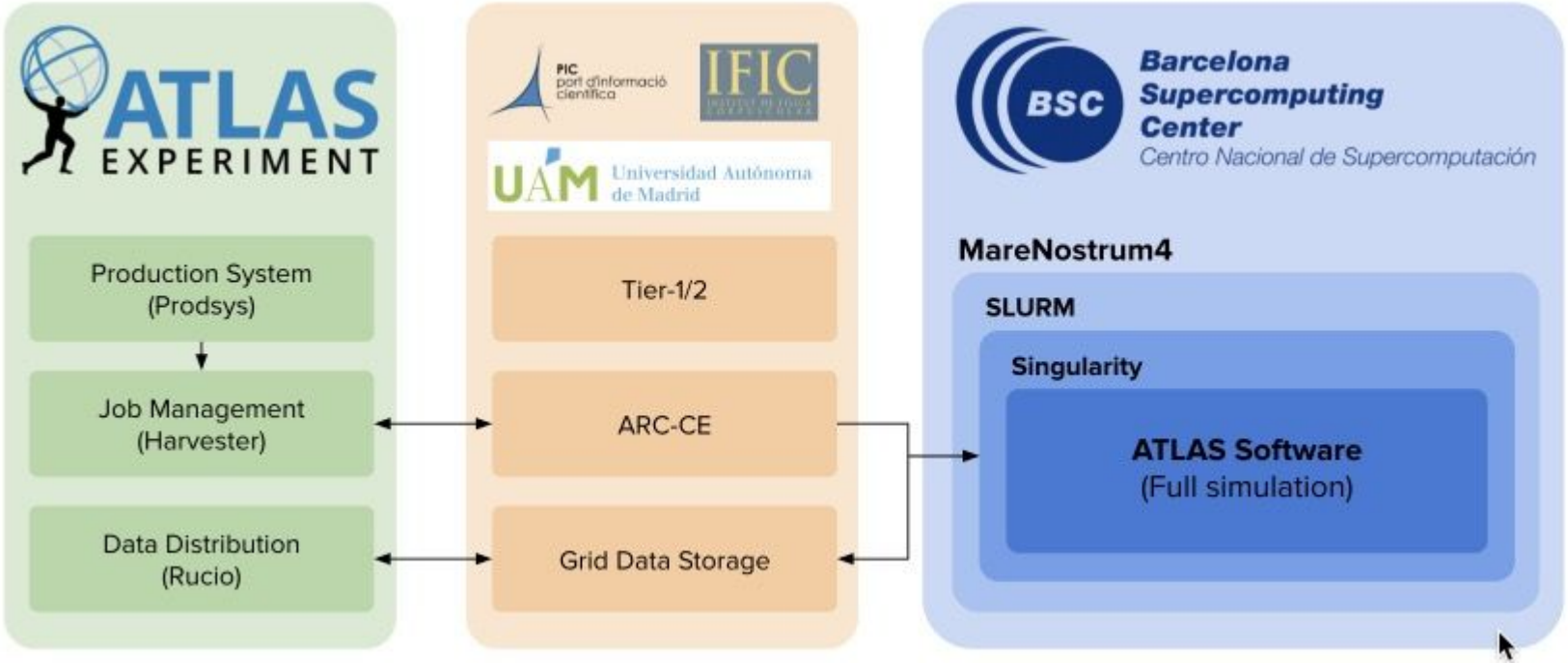
- We have opted for a flexible structure where the HPC center is one more resource that is added behind the existing Tier1 and Tier2 centers.
- Each center manages the requests for computing hours, and the total annual resources are negotiated as a single request.
- It is important for the Spanish LHC Community to always locate a Tier-1 or Tier-2 close to an LHC physics group.

Implementation using ARC-CE

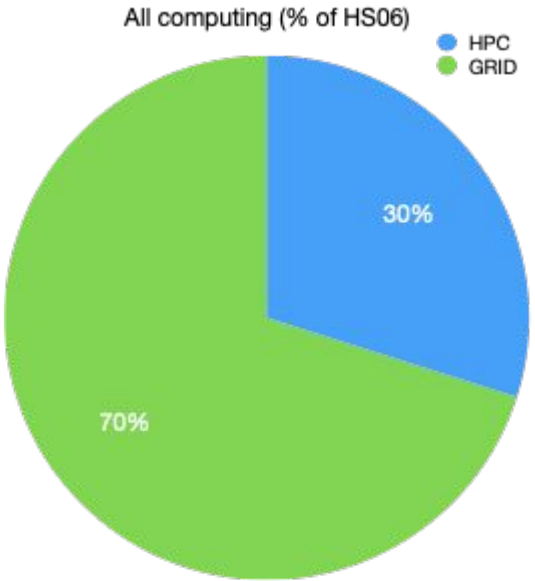
- Our main challenge is the lack of external connectivity at MareNostrum4 computing nodes.
- In Spain, we have used various approaches to get around this limitation.
- In this contribution, an ARC-CE, widely used in WLCG, has been used by ATLAS and LHCb. Harvester has been tested for ATLAS. And CMS has been testing a new HTCondor extension.
- In this presentation I will focus on the results obtained with ATLAS and the ARC-CE.

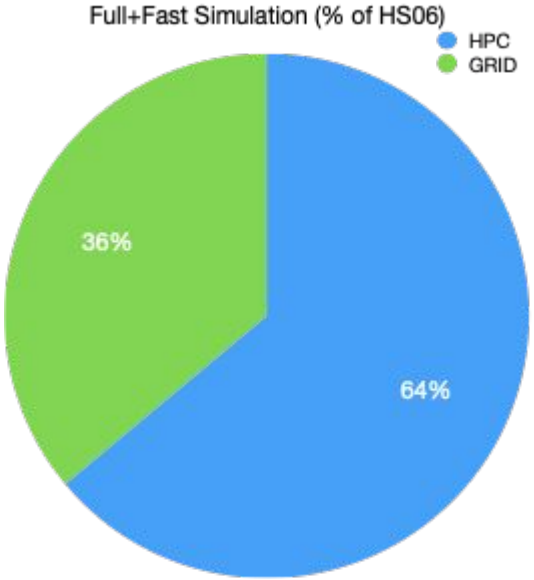
Workflow at MareNostrum 4

- We **copy all the input files** using the DTN by mounting a sshfs file system between the center and the BSC.
- We **submit the jobs** using the login nodes.
- The jobs run on validated **Singularity images** with all the **software and data preloaded**.
- We **check the status of the jobs** using the login nodes.
- We **retrieve the output files** using the sshfs filesystem.

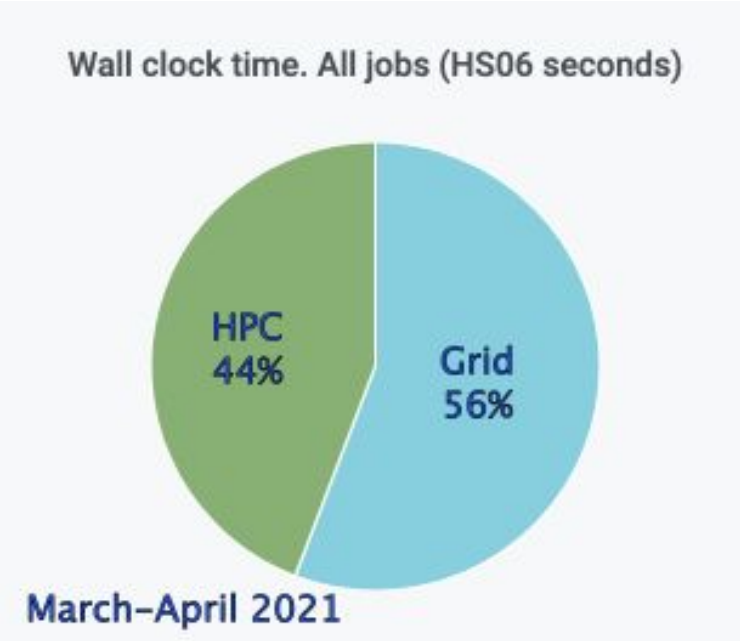

ATLAS EXPERIMENT
Production System (Prodsys)
Job Management (Harvester)
Data Distribution (Rucio)
PIC port d'informació científica
IFIC
UAM Universidad Autónoma de Madrid
Tier-1/2
ARC-CE
Grid Data Storage
BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación
MareNostrum4
SLURM
Singularity
ATLAS Software
(Full simulation)

Results from MN4 Integration: 2020





Results in 2021



Source: ATLAS Job Accounting

- On the left, we have the CPU consumption pie chart of ATLAS jobs by resource type in the last two months.
- We got 44% of the Spanish contribution to the CPU (all computing) from HPC.
- Target is to approach 50% for 2021.
- Only simulation is submitted to the HPC.

New developments

- Current plans are to increase the use of BSC thanks to the strategic program. We foresee to run 4 million hours per month.
- We will put effort
  - To increase the types of workflows we can run.
- After simulation the next target is the analysis jobs in containerized images.
  - Useful for analysis using GPUs

https://www.bsc.es/es/marenostrum/minotauro

Minotauro at BSC

- **100 nodes with Intel processors** E5649 (6-Core) or Intel Xeon E5-2630 (8-core)
- **2 GPUs per node** M2090 or K80 NVIDIA
- **4 GB RAM** per core
- Batch system: **SLURM**
- Operating system: **Red Hat Enterprise**
- Shared file system: **GPFS**

Can we replace the LHC computer centers?

- The answer is no.
- We need grid centers to receive the data from the experiment, store it on disk and **tape**, distribute, and **reprocess** the data. As well as to simulate and to analyze.
- The same is valid for simulated data once is produced, needs to be archived.
- The reconstruction of the data needs access to the databases of detector information, which is hard to upload to any supercomputer center.

Summary and conclusions

- We have managed to integrate the ATLAS Simulation jobs into the MareNostrum 4.
- The BSC has included the LHC computing in the list of strategic projects.
- We expect that the transition to MareNostrum 5 can be straightforward with 17 times more computing power in 2021.
- We still need grid computing for the LHC.
  - Still many workflows cannot run in the BSC due to the lack of connectivity.
  - We need to store, distribute and archive to tape the data.
- Thanks to the BSC and the Spanish Supercomputing Network (RES) for the resources.

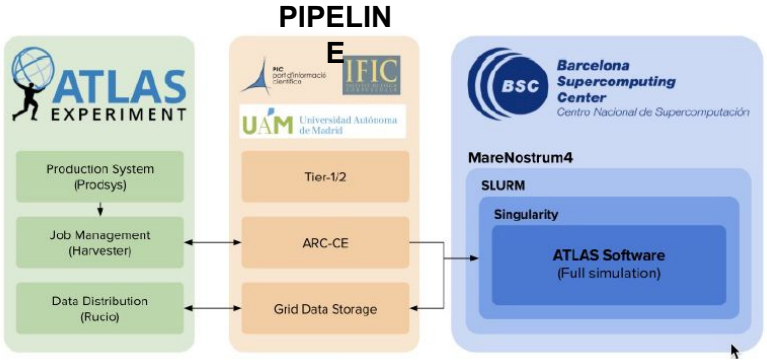

THANK
YOU

Backup slides

HPC ATLAS (IFIC, IFAE-PIC, UAM)

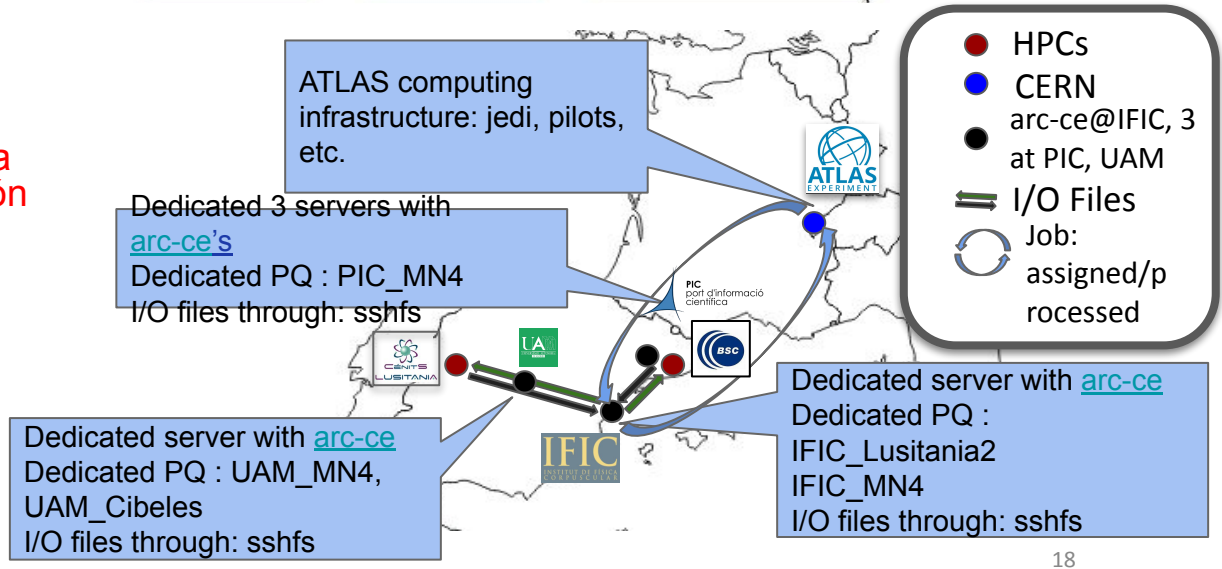
Use of HPC resources

- A large effort that is paying back
  - Started as an opportunistic resource **now it is a backbone of our computing contribution to simulation.**
- The access to HPC CPU time has been through the RES open calls.
  - From 2018 to mid 2020 as standard calls.
  - Starting in mid 2020 within the Ministerio-BSC agreement ("Proyecto Estratégico de Acceso al Marenostrum 4 para su utilización en la Computación del LHC").
- Three HPCs have been used Lusitania, Cibeles and MareNostrum4
- **LHCb** testing similar technical implementations in the same grant



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

- MareNostrum accepts only SSH protocol for job submission and data transfer

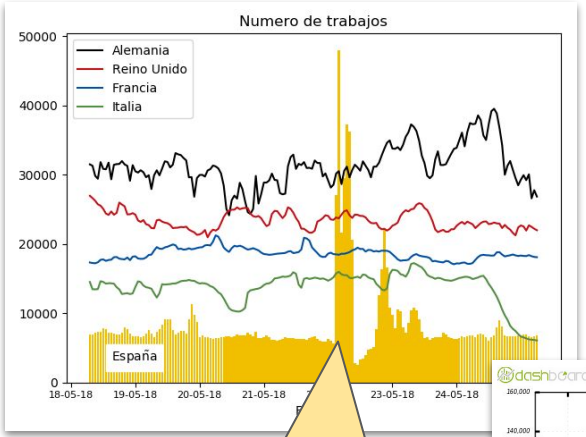


HPC ATLAS achievements

> 500k jobs processed

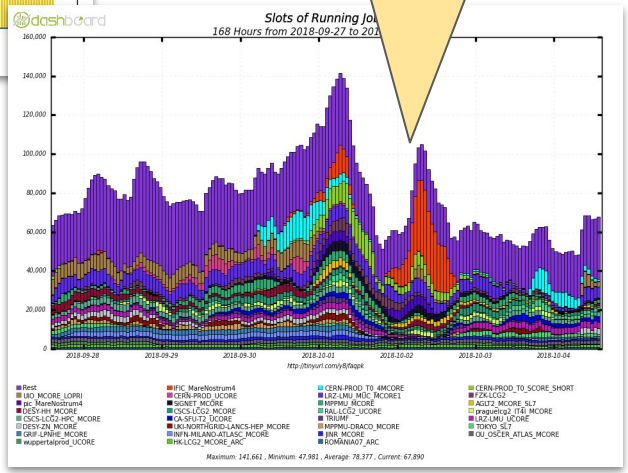
> 500M events simulated

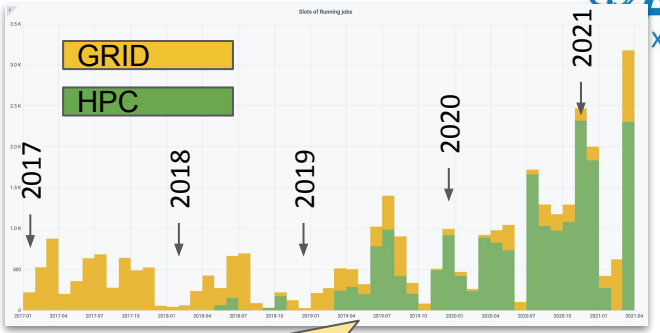
>30Mh CPU consumed



First week of HPC use by PIC and IFIC, Spain leads the ATLAS computing effort in Europe!

Number of slots running ATLAS jobs. In red Spanish Contribution (IFIC_MN4, PIC_MN4)





Evolution of slots of running jobs GRID-HPC from 2017 until now

Percentage of HS06 provided by GRID y MN4 since the agreement Ministerio-BSC. ONLY SIMULATION JOBS