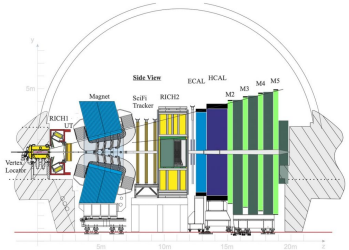
Computing in LHCb
WLCG and Use of BSC - MareNostrumV

Vanessa Acin (PIC), Carles Acosta (PIC), Alexandre Boyer (CERN), **Álvaro Fernández Casaní (IFIC/CSIC-UV)**, Pepe Flix (PIC), Jorge Lisa (IFIC/CSIC-UV), Fernando Martinez (IFIC/CSIC-UV), Xavier Vilasis (La Salle-URL)

XVII CPAN DAYS
Valencia 19/11/2025

LHCb data flow



4 TB/s
30 MHz non-empty pp

FULL DETECTOR READOUT

4 TB/s

PARTIAL DETECTOR RECONSTRUCTION & SELECTIONS (GPU HLT1)

0.5-1.5 MHz

70-200 GB/s

BUFFER

REAL-TIME ALIGNMENT & CALIBRATION

FULL DETECTOR RECONSTRUCTION & SELECTIONS (CPU HLT2)

10 GB/s

6% CALIB EVENTS

26% FULL EVENTS

68% TURBO EVENTS

1.6 GB/s

5.9 GB/s

2.5 GB/s

OFFLINE PROCESSING

ANALYSIS PRODUCTIONS & USER ANALYSIS

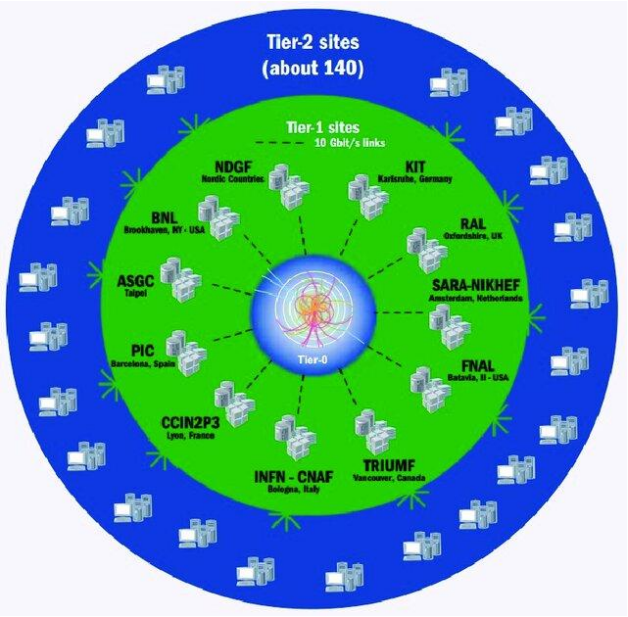
LHCb-FIGURE-2020-016

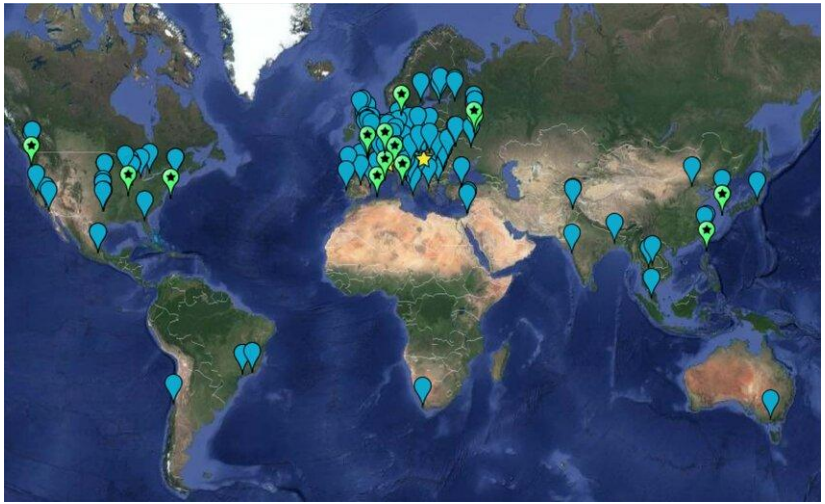
CERN HLT2 Cluster contribution

Installed March'25

- **56 x Computing Servers (x86_64):**
  - **AMD EPYC 9454P 48c/96t 2.75 GHz 290W**
  - **12 x 16 GB DDR5 = 192 GB**
    - 4 GiB/core (2 GiB/thread)
  - 1,92 TB SSD Enterprise (1DWPD/5 years).
  - 2 x Ethernet 10GBASE-T network.
  - 1 x 1000BASE-T port (BMC - IPMI)
  - RHEL 9.3 supported
  - **519,69 SPECrate (SPECrate2017_int_base) computing power, or**
  - **1888.3 HS23**
- **Total 2688 cores computing power = 29102,64 (SPECrate2017_int_base) or 105 kHS23**
- **Part of much bigger HLT2 cluster (>3k nodes)**
- **Used during online data taking, and offline use with DIRAC (see next)**

Distributed computing Grid

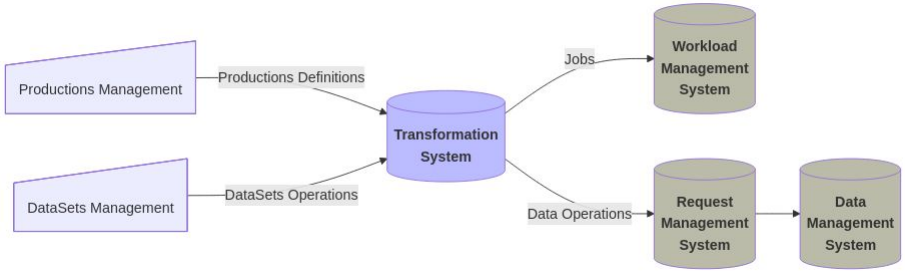




Distributed Resources Access

DIRAC THE INTERWARE → DIRAC^{X} THE INTERWARE

- **DIRAC** Manages jobs (pilots), productions, data management, productions, etc
- Started by **LHCb** now is used by several communities
- **WMS** uses pilots but also push jobs to HPC (see BSC MN5 approach next)
- **DM** own but be integrated with other tools, ie: Rucio
- **Transformation system** automates productions and data management
- **DiracX** (in development) is a complete rewrite of the software stack:
  - Cloud native, modular, multi-vo, standards based



Tier-1 Spain - Resources and Operations

- **LHCb Tier-1 contact (Álvaro Fernández)**
  - Operational issues and reports. Coordination with PIC T1 team and reporting, ie:
    - Software upgrades (arc-ce, dCache), Change tape paralelism RAW LHCb 2024, Incident on missing files and fixing catalog, new CEs on Alma9, etc.
- **Provided resources with LHCb MRR Funds managed by IFIC:**
  - **2 x Disk Servers with net capacity: 960 TB**
    - 2 x Procesador Intel 4314 2P 16C/32T 2.4G 24M 10.4GT 135W
    - 16 x 32GB de memoria DDR4-3200 2Rx8 (Total: 512)..
    - 24 x Disk Seagate 3.5",24TB,7.2K RPM,SAS3 12Gb/s,512e/4Kn (Summit).
    - 2 x SFP28 Transceiver module 25G, 850nm, MMF, LC.
  - **(To be deployed) 9 x Disk Servers with net capacity: 4320 TB**
    - Intel Xeon Gold 6426Y (16C/32T, DDR5-4800, 185W)
    - 512 GB DDR5 (8×64GB - 5600MT/s)
    - 24 x Disk SAS ISE de 24 TB 7.2K 12Gbps (raw: 576 TB / net 480TB). RAID 6
    - 2 x SFP28 SR Optic, 25GbE, 85C

Installed Feb'25

Early 2026

Tier-1 Spain Pledges

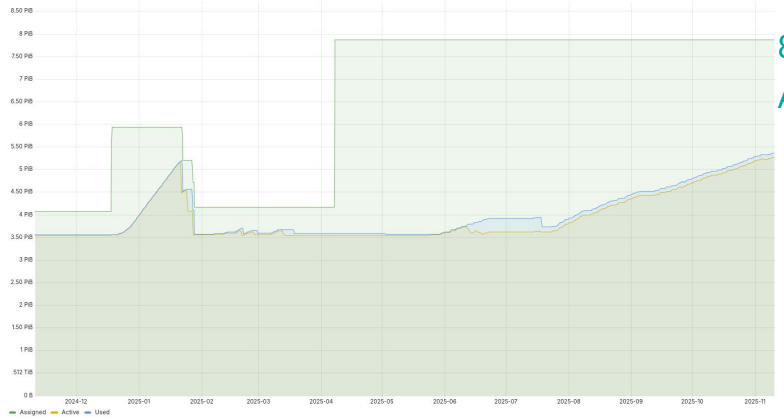
- Tier-1 Pledges (from P.Flix/PIC)

		2023	2024	2025	2026
ATLAS T1	CPU (kHS23)	57.200	60.640	73.575	90.100
	Disk (PB)	5.440	6.540	8.395	9.950
	Tape (PB)	14.120	18.080	25.245	34.600
CMS T1	CPU (kHS23)	32.000	37.200	49.500	60.000
	Disk (PB)	3.920	4.880	6.390	8.200
	Tape (PB)	12.640	15.200	20.025	27.000
LHCb T1	CPU (kHS23)	28.280	22.880	41.760	56.350
	Disk (PB)	2.420	2.448	4.046	5.355
	Tape (PB)	6.280	5.332	8.766	11.685
	WLCG T1 share	4%	4%	4.5%	5.00%

(*) LHCb estimates Spain should contribute ~8%

Tier-1 Spain LHCb - Tape and Disk Resources

Tape



8.86 PB
Assigned

6.04 PB
Used

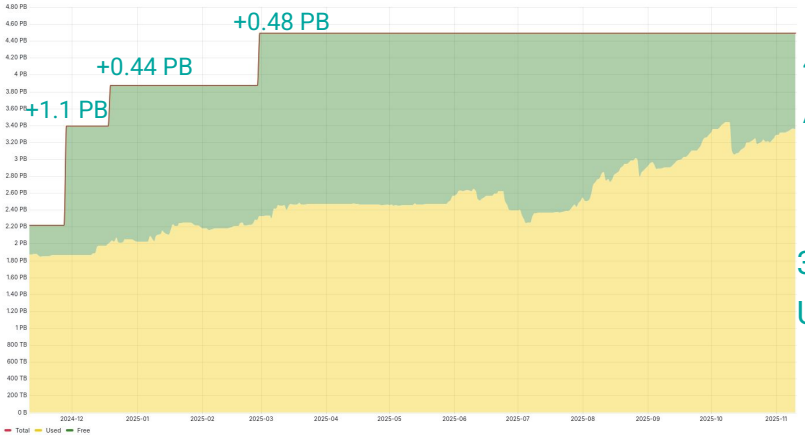
LHCb Tape :

Assigned: 8.86 PB (Pledge 2025: 8.766 PB)

Used: 6.04 PB

Images showing last 1 year (Nov2024-Nov2025)

Disk



4.67 PB
Assigned

3.42 PB
Used

LHCb Disk:

Assigned: 4.67 PB (Pledge 2025: 4.04 PB)

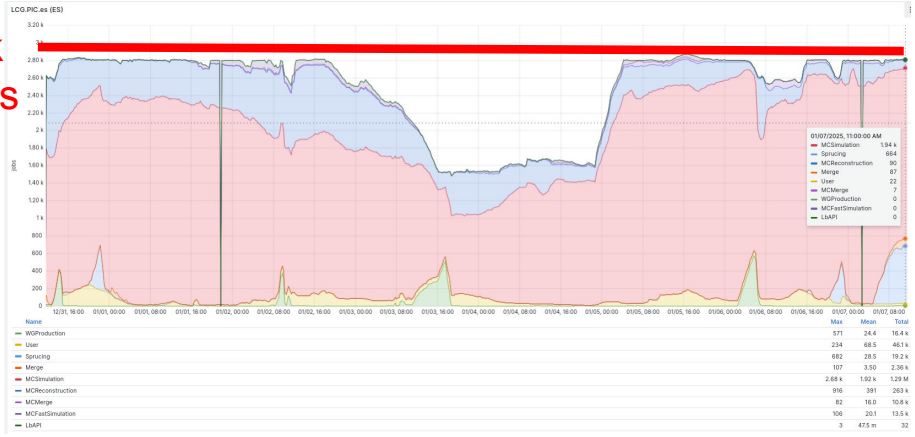
Used: 3.72 PB

PIC added Disk servers during last months.

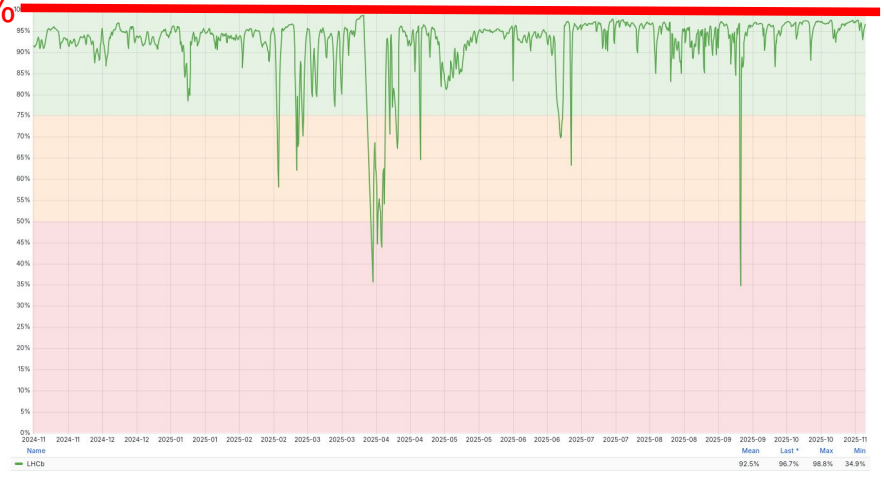
Feb'25: +0.48 PB of the 0.96 PB (2 disk servers - IFIC MRR Funds)

Tier-1 Spain LHCb - CPU Resources

3k jobs

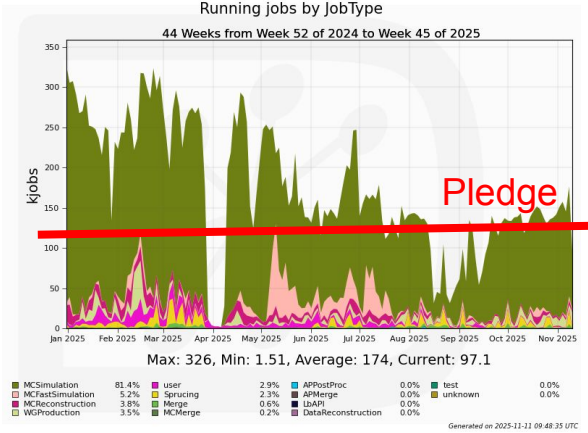


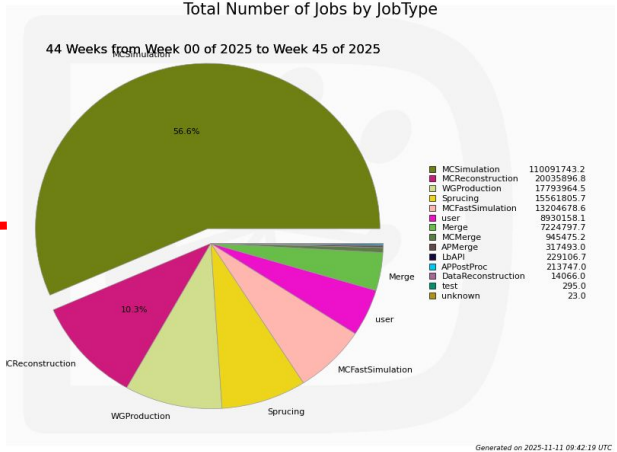
100% Eff

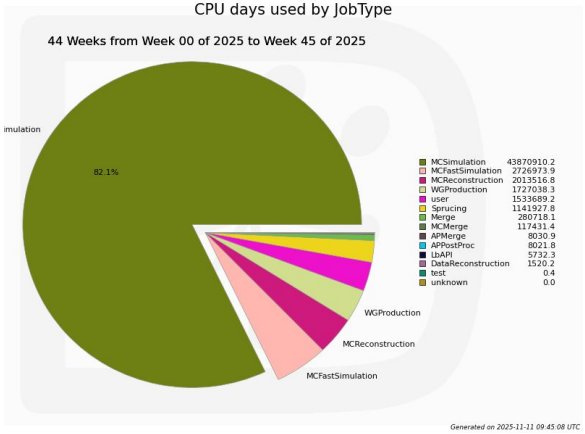


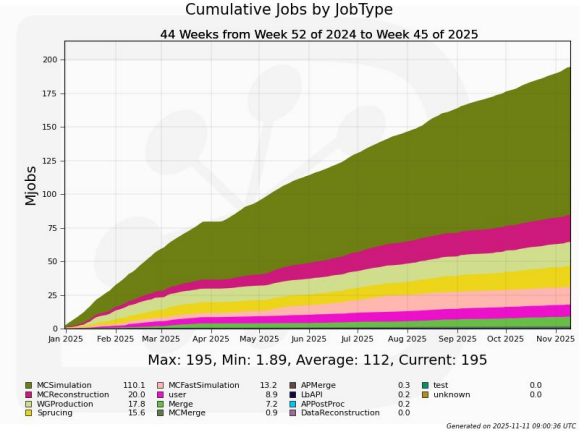
- Usual 3 k jobs fair share usage among all queues
- MonteCarlo simulation takes most of the resources
- High CPU efficient single-core jobs

LHCb - Worldwide CPU Resource Usage 2025



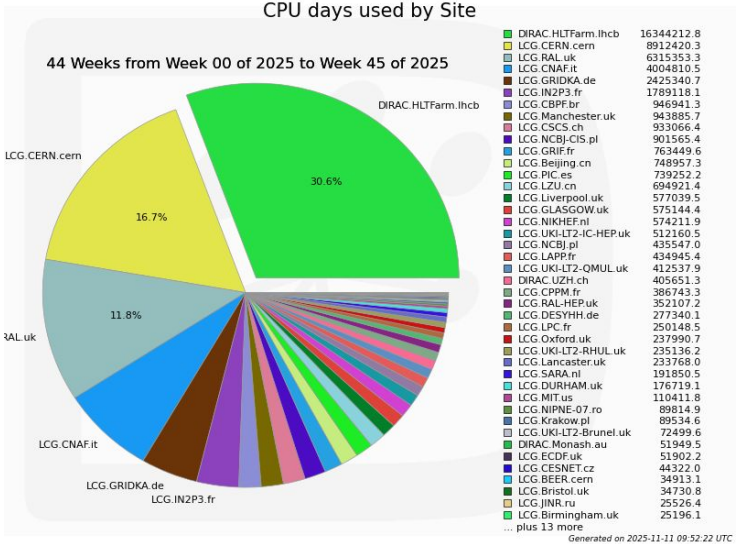


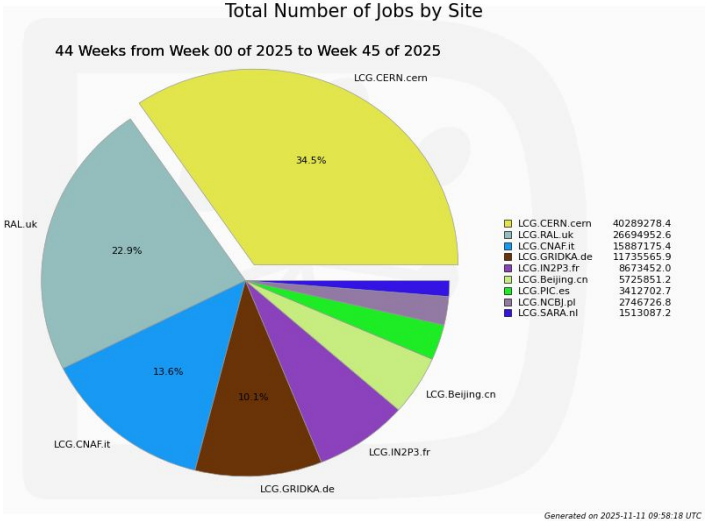




- **195 Million Jobs** were executed
- Majority are **MonterCarlo simulations** (80% of jobs, **90% of CPU time**)
- Pledge was **120k jobs** based on an average HS23 score over the year

LHCb - Worldwide CPU Resource Usage by Site 2025



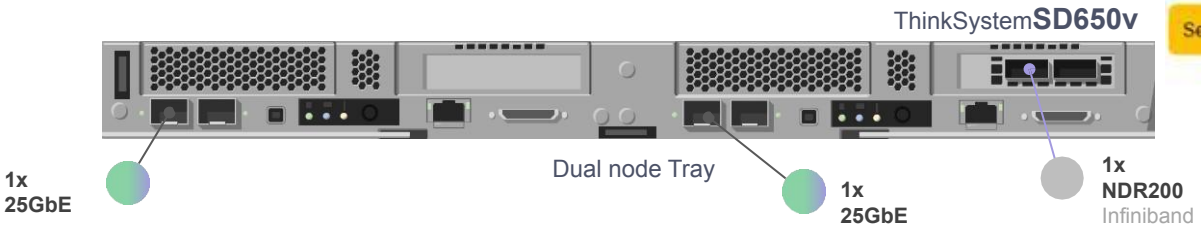


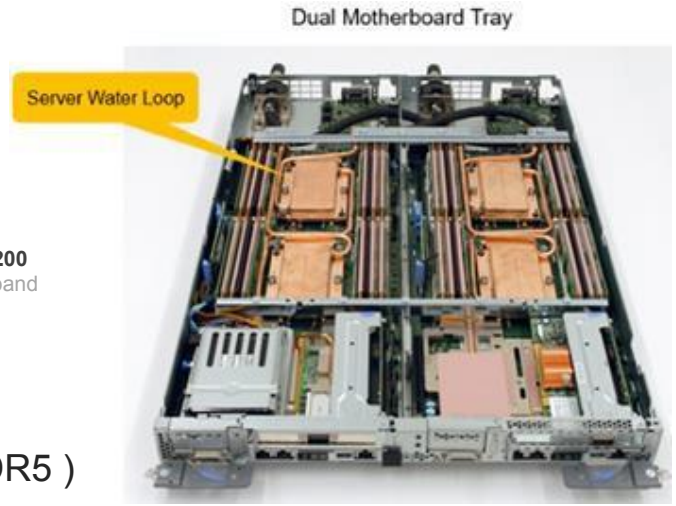
- **HLT farm** (used when not data taking) provides **30% of CPU time** (21% of jobs)
- **CERN TIER-0** provides **16% of CPU time** (20% of jobs)
- **TIER-1 sites** provide **31% of CPU time** (39% of jobs)
  - **PIC** 6th with 1.38% of CPU time (1.75% of jobs)
- **Tier-2** sites provide **~21% of CPU time** (18% of jobs)
- Rest provided by opportunistic/non-pledged resources

BSC MareNostrum 5

- **Barcelona Supercomputing Centre (BSC) new MareNostrum5 cluster** since April 2024.
- **314 Petaflops** peak performance
- **223 M€** of investment
- **MareNostrum 5 GPP (General Purpose Partition)**: 6480 (CPU) nodes with total 725,760 processor cores and 1.82TB of main memory. 45.4 Pflops
- **MareNostrum 5 ACC (Accelerated Partition):** 1,120 nodes based on Intel Xeon Sapphire Rapids processors and 4xNVIDIA Hopper H100 GPUs. 260 Pflops
- **MareNostrum 5 NG-GRACE:** 408 nodes, each powered by NVIDIA's Grace CPU Superchip (ARM).

General Purpose Compute Node





6,192x GPP Compute node (256GB RAM 16x16 GB 4800MHz DDR5)

216x GPP Compute node (1TB RAM 16x64 GB 4800MHz DDR5)

72x GPP HBM Compute node (32GB RAM 2x16 GB + 128 GB HBM2)

- 2x Intel Xeon Platinum 8480+ 56C 2GHz (**112 cores)**
- 16x DIMM 16GB/64GB 4800MHz DDR5 (**256 GB**)
- 960GB NVMe local storage
- ConnectX-7 NDR200 InfiniBand (shared by two nodes, 100Gb/s bandwidth per node)

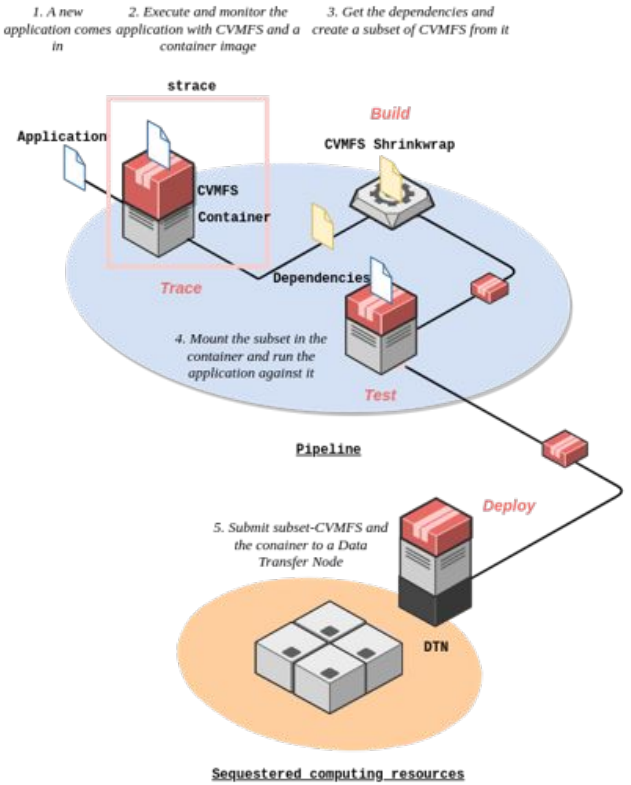
Usage and Limitations

- **Public Login Nodes** give access to the infrastructure.
  - only nodes accessible from external networks.
  - No outbound connectivity.
  - SSH access only.
- **No custom software installation possible**
  - RHEL 9.2 Operating System
  - SLURM resource manager
  - Singularity and other standard software available configurable with Lmod.
- **Limit on number of jobs (unicore/multicore/multinode) per account in the system: 366 jobs/account**
- **Scheduling (in principle) favors complete-node allocations (ie: 112 cores)**
- **Usable memory per core:**
  - **2 GB/core in GPP nodes**
  - 9 GB/core in GPP-HImem nodes (limited)

- **Pilot jobs (pull model) not possible**
  - We use another approach to push jobs that execute without network requirements
- **CVMFS software distribution not possible**
  - We require a custom software distribution method
- **Multicore jobs are preferred over single core jobs**
  - Our current workload is single core (Gauss)
  - We would need a multicore approach (next MC simulator Gaussino)
  - Or an approach to Bundle multiple single-core instances into a multi-core job

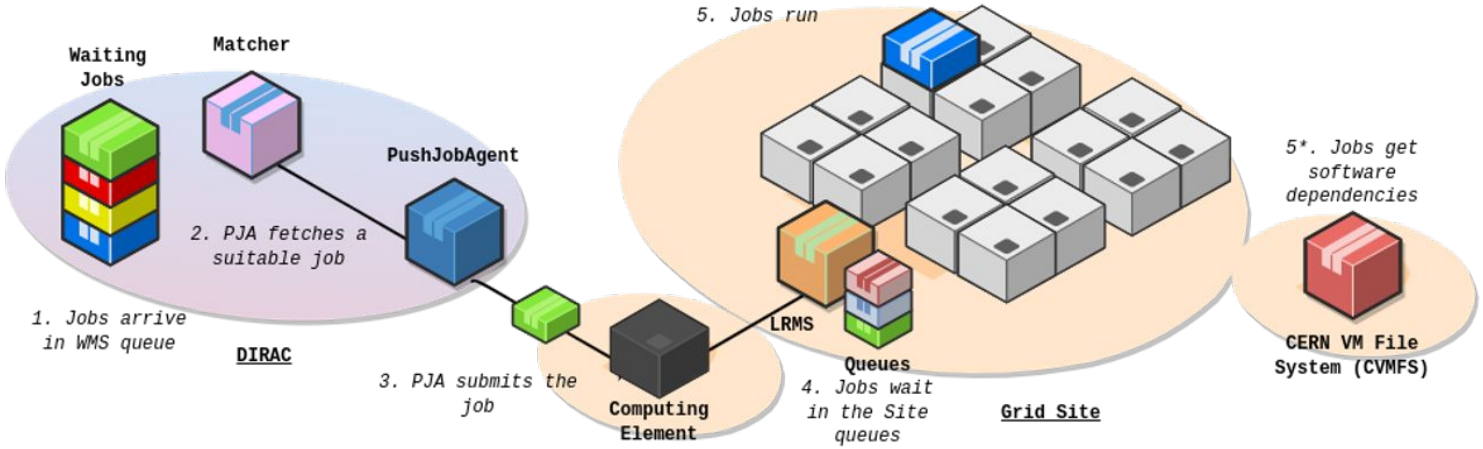
Software Distribution - CVMFS snapshot

- **CVMFS not mounted on nodes**
- **We have limited storage space available in MN5 LHCb project**
- **Solution: a snapshot of CVMFS copied to MN5 storage**
- **Automated with SubCVMFS Builder**
  - **Continuous integration pipeline** running at GitLab CERN.
  - Runs the target software (i.e. Gauss) inside container to **trace needed files and packages**.
  - Include the required Dirac(x) distribution.
  - **Builds only if necessary** (new software version or dependencies) and production the snapshot (~30GB of dependencies for Gauss)
  - **Copies only incremental changes** (rsync)



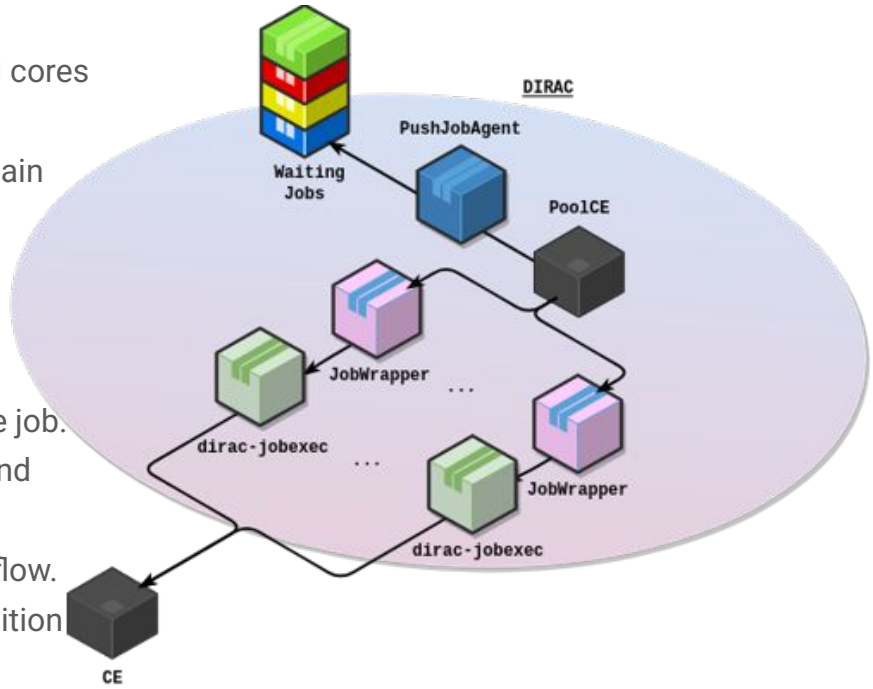
Job Submission - Push Model

- **Lack of external connectivity** does not allow pilot jobs (pull model)
- **Solution: PushJobAgent (push model)** Fetches jobs, manages their input and output data, and solely submits the application (Gauss) to MN5
- **Main Limitation:** Every job running at MN5 requires a memory consuming 'shadow' process at DIRAC vobox servers, limiting scalability



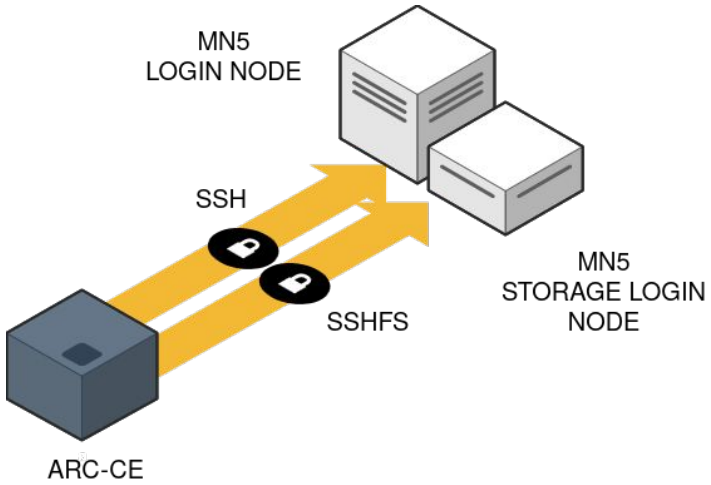
Overcoming Limitations

- **Gaussino (LHCb simulation team):** multi-threaded MC Simulation application, intended to replace Gauss.
  - Expected soon, this would potentially improve x 112 (all cores full-node) usage
  - Correctly tested a pre-release version within all Dirac chain
- **New PushJobAgent** able to submit complete jobs.
  - Stateless and more robust
  - More scalable: memory consumption O(1)
- **New BundleCE and LHCb jobs.** ONGOING
  - To be able to bundle several single-core jobs in a unique job.
  - Only keep the Gauss module in the "DIRAC workflow", and make it completely offline.
  - Other modules and steps are handled outside the workflow.
  - Using CWL (Common Workflow Language) for job definition

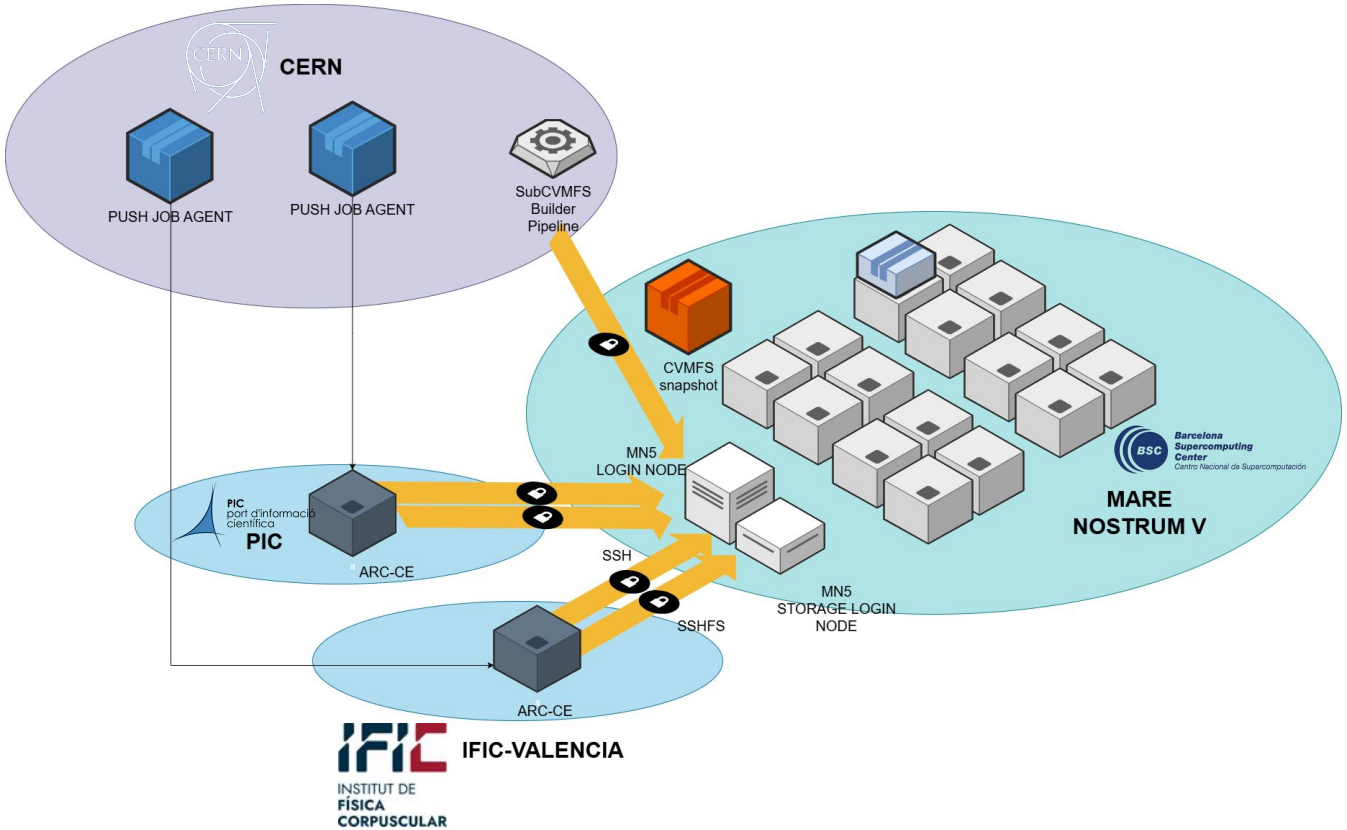


ARC-CE Endpoint

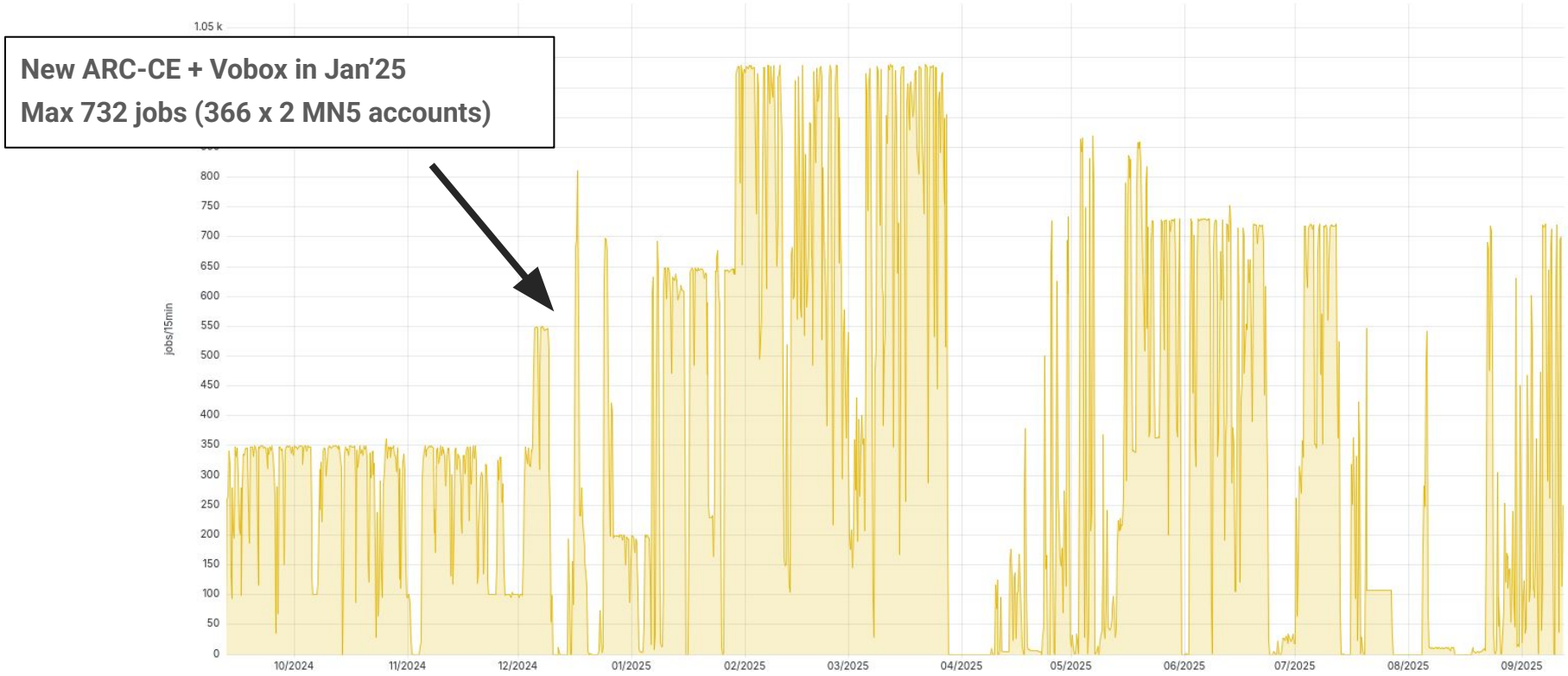
- **PushJobAgent contacts ARC-CE as gateway to MN5**
  - Submits input files, monitors jobs execution, retrieves output files.
- **ARC-CE 7**
  - Configured to submit to SLURM as “local” resources
    - Modified slurm commands execute remotely with SSH to the MN5 Public Login Node
  - SSH Passwordless Public key authentication mapping single MN5 authorized user account.
  - Mounts MN5 remote storage with Fuse SSHFS to share ARC SessionDir.
  - Allows submission with OIDC tokens instead of x509 cert.
- **Singularity Run Time Environment**
  - Configured for the job to start running in MN5 a Singularity image that mounts the CVMFS snapshot, and execute the job payload



Complete Picture - Current Deployment

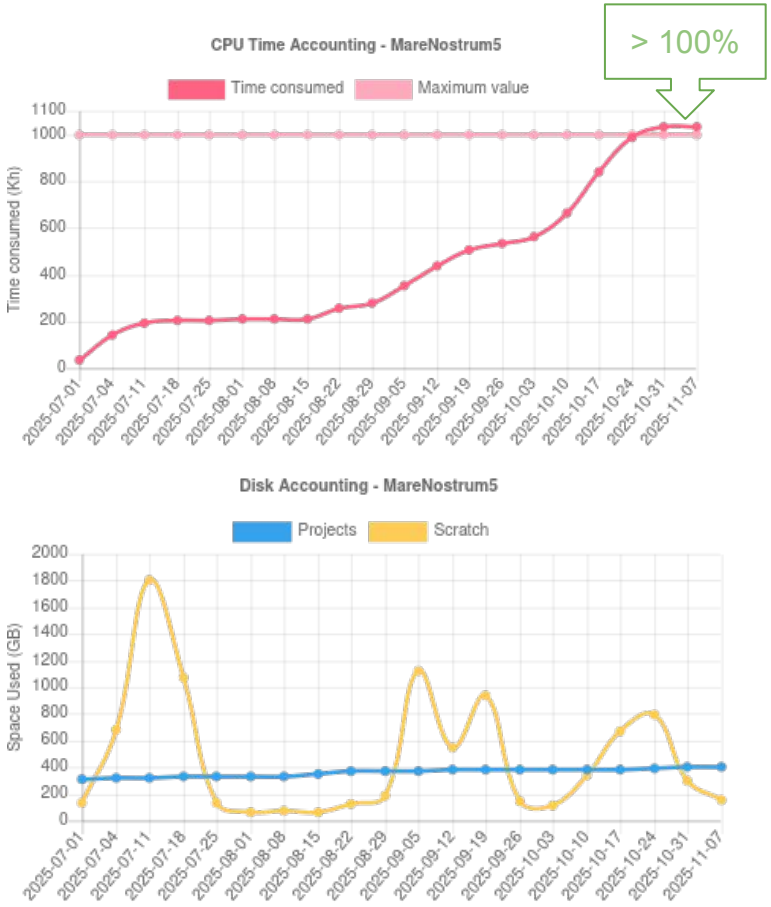


Number of LHCb jobs @ BSC MareNostrum 5



LHCb @ BSC MareNostrum 5

- **A Call submitted every 6 months now (was every 4 month until September 2025)**
- **Doubled the number of requested CPU hours since last year (1M hours last call)**
  - More powerful CPU in MN5 compared to MN4 also means more delivered simulations.
- **Current Disk allocation:**
  - **4 TB projects (CVMFS Snapshot)**
  - **4 TB scratch (Jobs SessionDir)**
- **To request more CPU hours**
  - Lack of outbound network connectivity forces to use push model.
  - single-core jobs (Gauss) -> multicore Gaussino
  - Improvements on scalability of current solution



Conclusions

- **LHCb distributed computing using HLT2 farm, grid (WLCG) and HPC supercomputer resources.**
- **LHCb MRR Funds managed by IFIC** contributed to
  - CERN HLT2 cluster (online data taking, and offline use with DIRAC)
  - Tier-1 storage resources
- **LHCb exploits MareNostrum 5 reasonably well**
  - Lack of outbound networking and no software installation limits working model.
  - Overcoming the MN5 limitations with the PushJobAgent.
  - 1 Million CPU hours each 4-month period (now 6-months), to be increased.
  - New ARC-CE at IFIC-Valencia allowed multiply x2 number of submitted jobs.
- **Other improvements are in progress**
  - **New PushJobAgent approach and New LHCb Jobs:** using CWL (Common Workflow Language) and new approach to increase scalability and reduce operational tasks
  - **Multi-core Gaussino** reaches production: x112 (full-node) usage