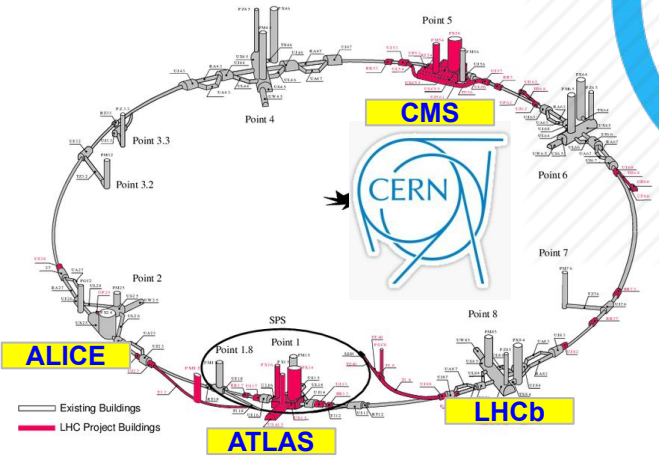
Enabling HPC systems for HEP: The INFN-CINECA Experience

Stefano Dal Pra (INFN-CNAF)
on behalf of many people, see last slide

PRACE

WLCG
Worldwide LHC Computing Grid



The 3 parties in the game

- LHC Experiments
  - The italian groups in ALICE, ATLAS, CMS, LHCb, deliver 10-15% of the experiments' computing pledges.
- CNAF
  - Italian Tier-1 in Bologna. Main INFN data center providing computing and storage to more than 30 experiments
- CINECA
  - HPC center in Bologna. EU/Prace Tier-0.
  - In 2019 top machine was **Marconi**, partially **KNL** and partially **SkyLake** based. At **21st position in top500.org**
  - Selected site for 1 of the 3 EU pre-exascale machines

CNAF

- Standard "GRID-like"HTC farm (30k cores, 400 kHS06)
- 38 PB of disk
- 90 PB of tapes on 2 libraries

CINECA

Marconi cluster

- Based on Omnipath
- ~19 Pflop/s
- 17 PB of local storage

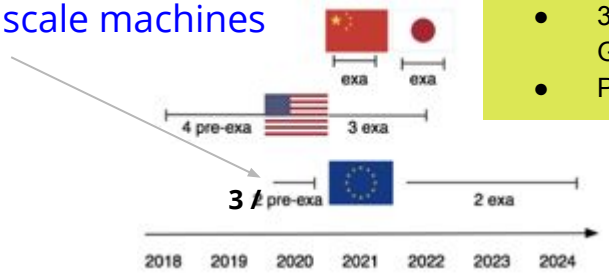
Marconi A2 Partition

- 3600 nodes with 1 Xeon Phi 2750 (KNL) at 1.4 GHz and 96 GB of RAM
- 68 cores/node, 244800 cores
- Peak Performance: ~11 Pflop/s

Marconi A3 Partition

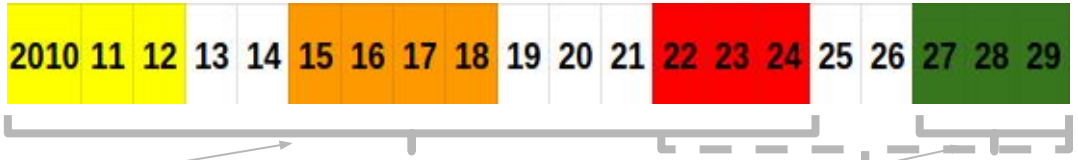
- 3216 nodes with SkyLake at 2.1 GHz
- Peak Performance: ~8 Pflop/s

Rank	Site	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)
21	CINECA Italy	Marconi Intel Xeon Phi - CINECA Cluster, Lenovo SD530/S720AP,	348,000	10,384.9	18,816.0

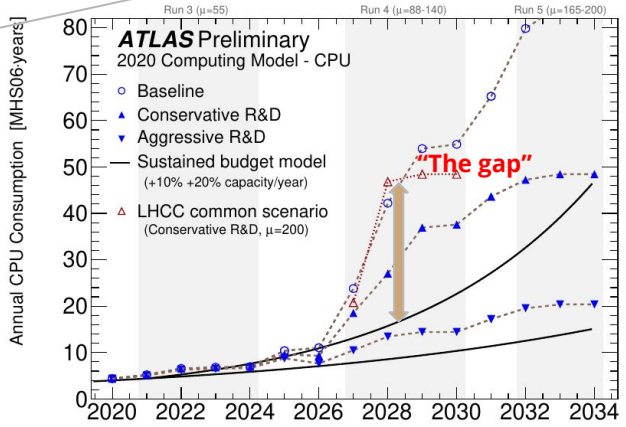


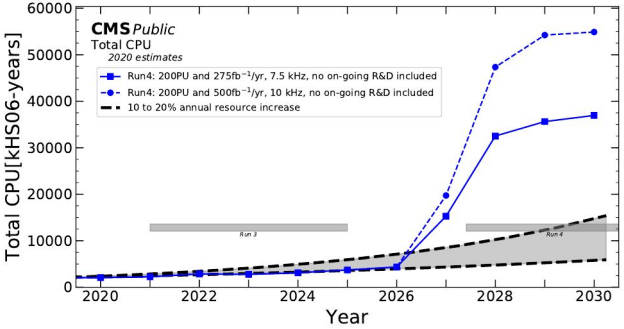
A typical Marconi A2 node was configured with	A typical WLCG node has
A KNL CPU: 68 or 272(HT) cores, x86_64, rated at ~¼ the HS06 of a typical Xeon	1/2 Xeon-level x86_64 CPUs: typically 32-64 cores, O(10 HS06/thread) with HT on
96 GB RAM, with ~10 to be reserved for the OS: 1.3-0.3 GB/thread	2GB/thread, even if setups with 3 or 4 are more and more typical
No external connectivity	Full outgoing external connectivity, with sw accessed via CVMFS mounts
No local disk (large scratch areas via GPFS/Omnipath)	O(20 GB/thread) local scratch space
Access to batch nodes via SLURM; Only Whole nodes can be provisioned, with 24 h lease time	Access via a CE. Single thread and 8 thread slots are the most typical; 48+ hours lease time
Access granted to individuals	Access via pilots and late binding; VOMS AAI for end-user access

Does it make sense to try an integration?



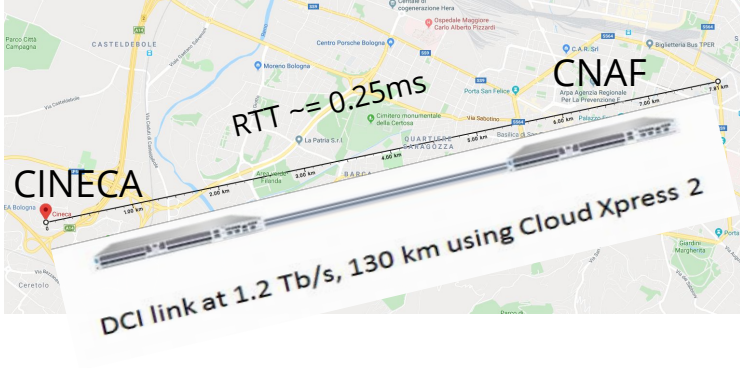
- Today
  - we are not relying on HPC resources. **Not on the critical path**
- Tomorrow (2027+)
  - There are strong hints that HL-LHC processing will need to access HPC resources depending on **specific Funding Agency's policies**
  - The current modelling of HL-LHC computing does not fit a reasonable budget with only in-house resources
  - We need to build experience on today's HPC systems in order to be prepared and influence next generation systems
  - HPC systems give massive access to accelerators and in general to heterogeneous systems - we need to be prepared
- On top of that ...

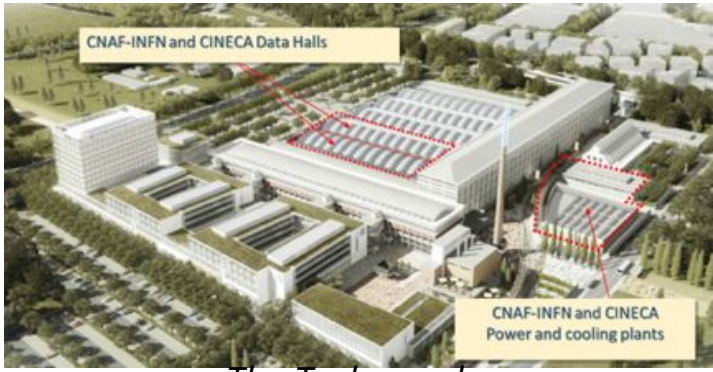




CNAF and CINECA

- **Close** to each other (less than 8 km)
- Existing collaboration: ~50% of the CNAF CPU power (180 kHS06, 2018) already on regular “GRID-like” nodes hosted @ CINECA (refurbished from the former A1 partition of Marconi)
- Connected via a dark fiber and a pair of Infinera CloudXpress systems
  - Capping at 1.2 Tbit/s; currently at half capacity
  - High bandwidth + low latency → no strict need for caches
- CINECA and CNAF are planning the migration to a common location (the “Technopole”) by end of 2022, and integration experience is welcome in that perspective
- **CINECA and INFN are partners in the Pre-Exascale Leonardo system**





The Technopole

The Grant

- The LHC Italy community successfully applied for a **“PRACE Project Access”** on the CINECA KNL partition
- **30McoreH allocated** after a demonstration the project was “feasible” (via a 20kcoreH test)
- “**Feasible**” ~~meant many~~ handshaking / changes to initial setup on both sides - thanks to the mutual understanding and the flexibility from CINECA’s side on what is seen as a use case of mutual interest

Project scope and plan

Project name	LHC@HPC
Research field	High Energy Physics

Principal Investigator (PI)

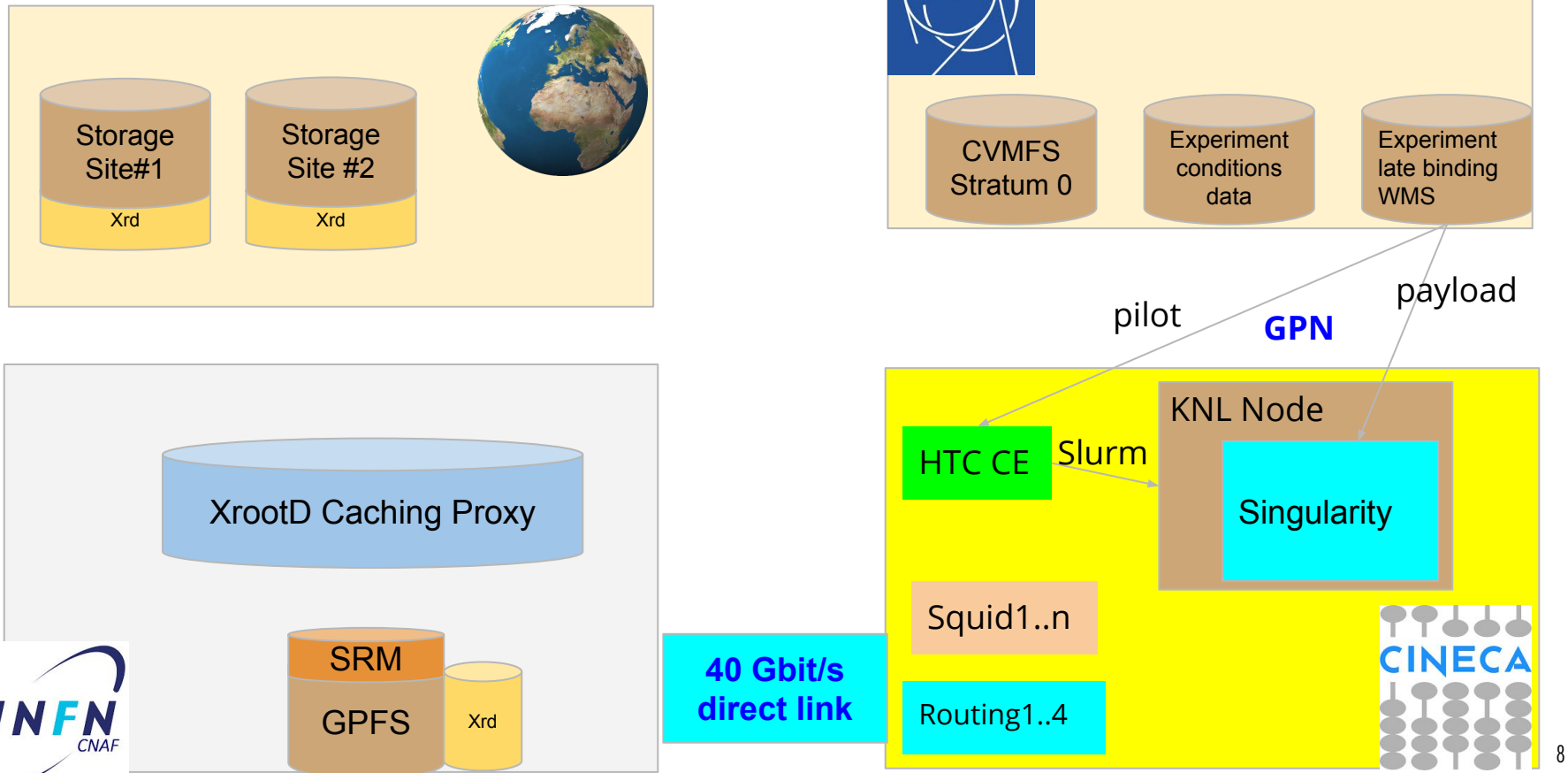
Title (Dr., Prof., etc.)	Dr
Last name	Boccali
First name	Tommaso
Organisation name*	Istituto Nazionale di Fisica Nucleare (INFN)
Department*	Sezione di Pisa
Group*	Scientific Computing, CMS
Country	Italy

- **CVMFS** (and its squids) **ok**
  - Interest also from non HEP community
- **External networking** enabled to **CNAF and CERN**
  - Enough to guarantee access to CNAF storage and conditions for experiments’ workflows
  - We can use it also for accessing external data
- Partial routing to CNAF storage / squids over the dark fiber @ **40 Gbit/s** (technical limitation to be removed with next machine)
  - The rest over GPN
- **Singularity** audited by CINECA’s sysadmins and green lighted
- A **HTCondorCE/Slurm** allowed at the CINECA edge nodes

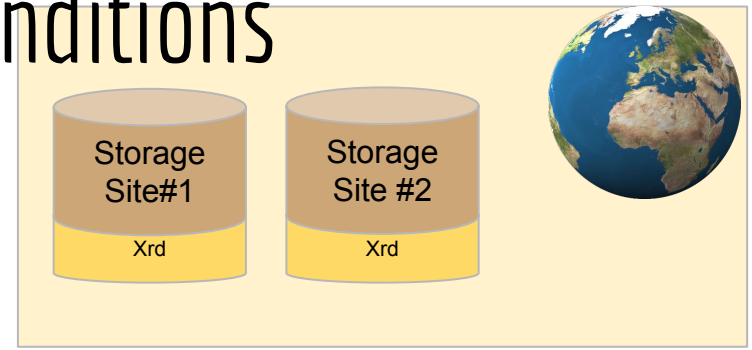
Some personal comments:

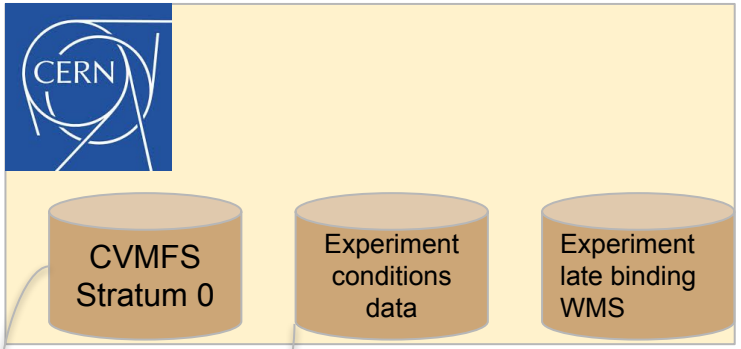
- Being able to speak to the center makes things so much easier
- Often the site limitations are there since “nobody needed the feature before”
- In general, getting an agreement on the changes was easier than feared!

Technical setup#1: jobs

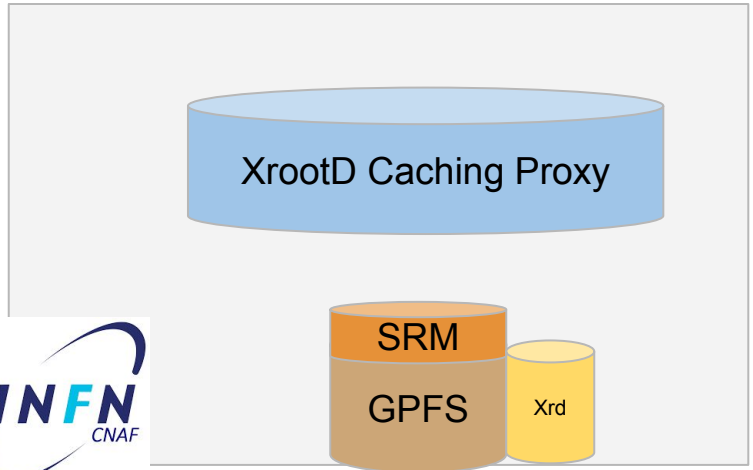


Technical setup#2: sw and conditions

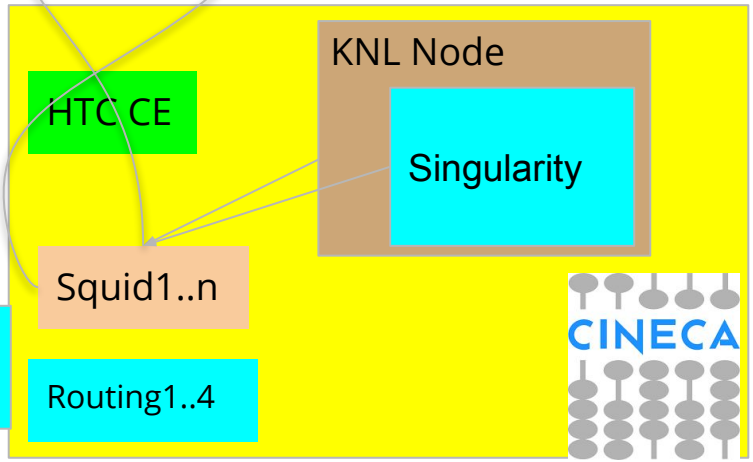




GPN

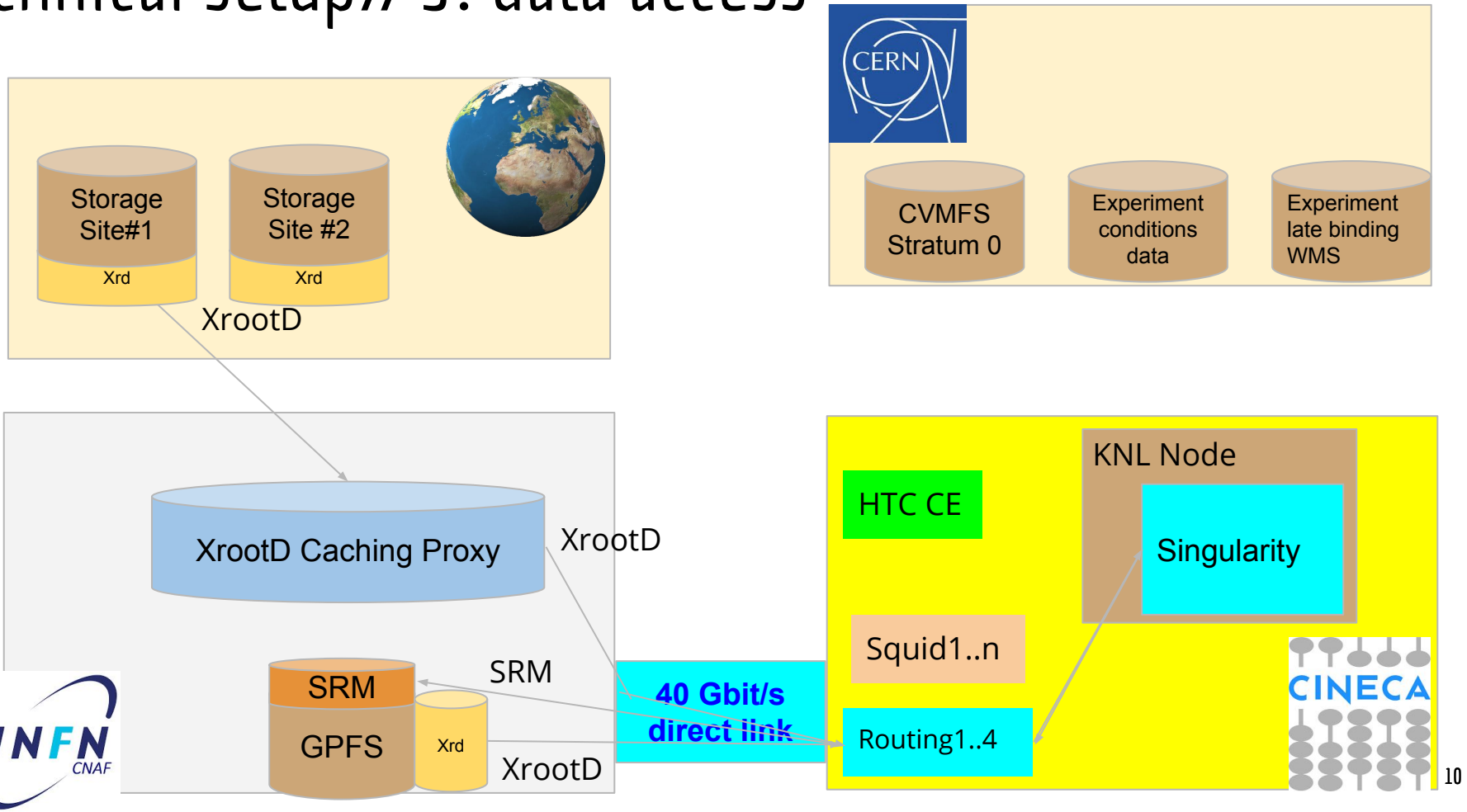


40 Gbit/s direct link



INFN CNAF

Technical setup#3: data access



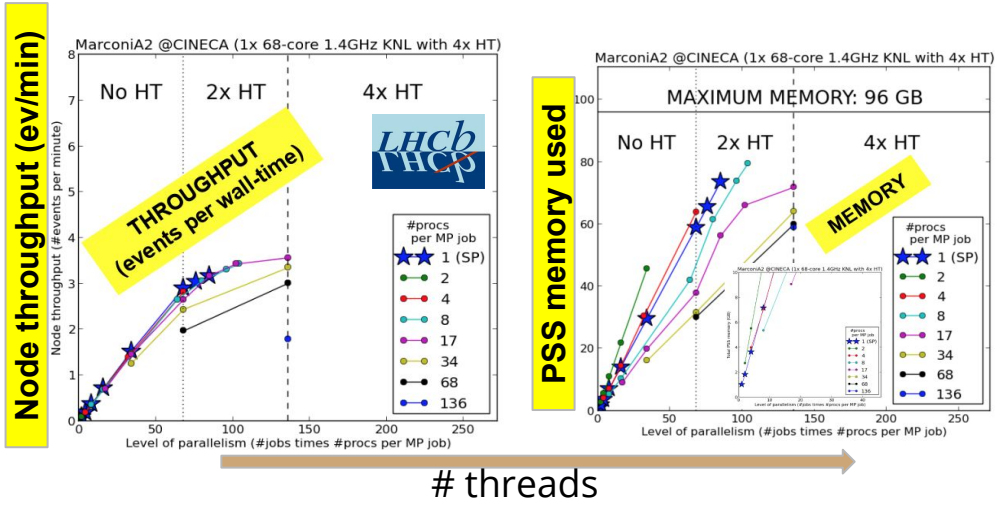
… but: what to run?

- Nodes with 272 threads each (68 physical cores with 4x HT)
  - Insufficient RAM (86+10 GB) to run 272 single threaded payloads
- Clear advantage in using MT/MP codes to reduce the footprint
  - Need to find the best configuration, different from the one used on WLCG machines with 2 GB RAM / thread
  - Effort needed on software application development to implement MP or MT
  - Effort needed on Grid job submission to fill as many of the 272 threads as possible using many MT/MP jobs

Examples from the four LHC experiments on the next slides!

LHCb

(F. Stagni et al. @CHEP 2019) arxiv:2006.13603

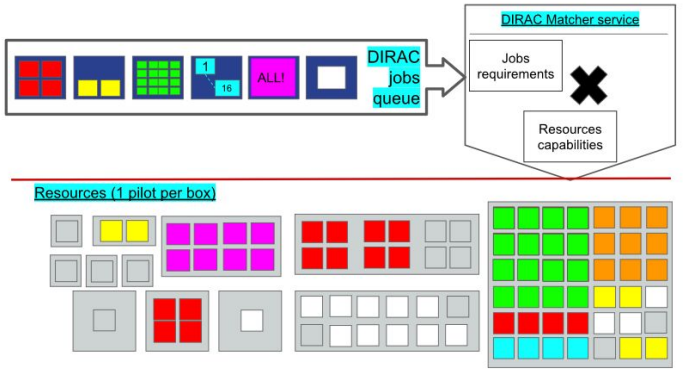


LHCb MC Simulation software

- RAM enough for ~80 single-process jobs
- Using a multi-process application, may use 136 threads with 8 MPx17 jobs, but with minimal gain in event/sec throughput

LHCb multi-core Grid submission

- DIRAC Matcher service
- One pilot per node, partitions the node for optimal job allocation



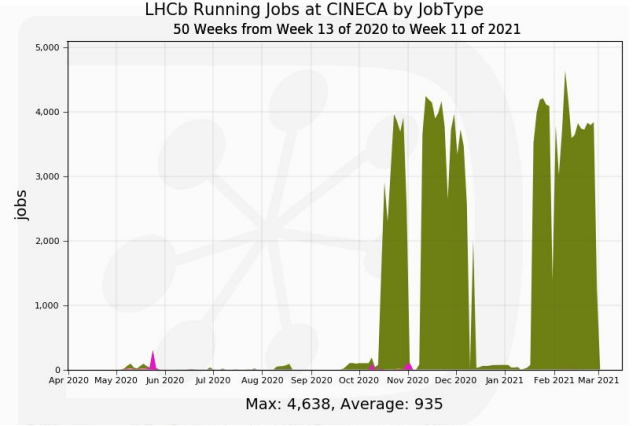
LHCb usage of CINECA

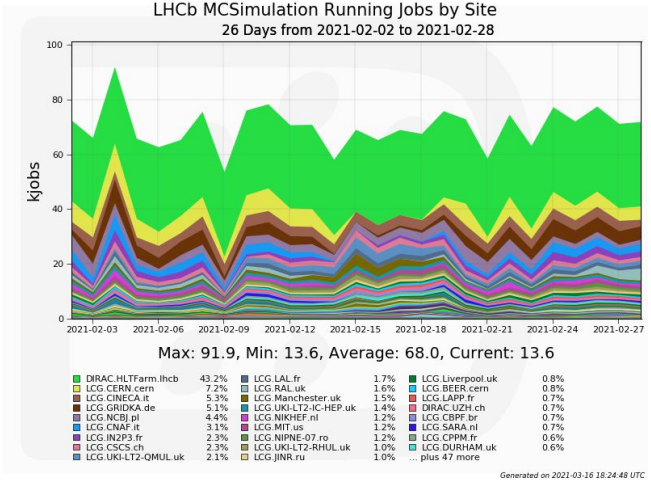
Detailed MC simulation: CPU intensive, no input data, output data to CNAF

LHCb pursued and achieved two goals

First goal (Q2 - Q3 2020): commission multi-core job submission via Dirac in production (multi-process MC jobs)

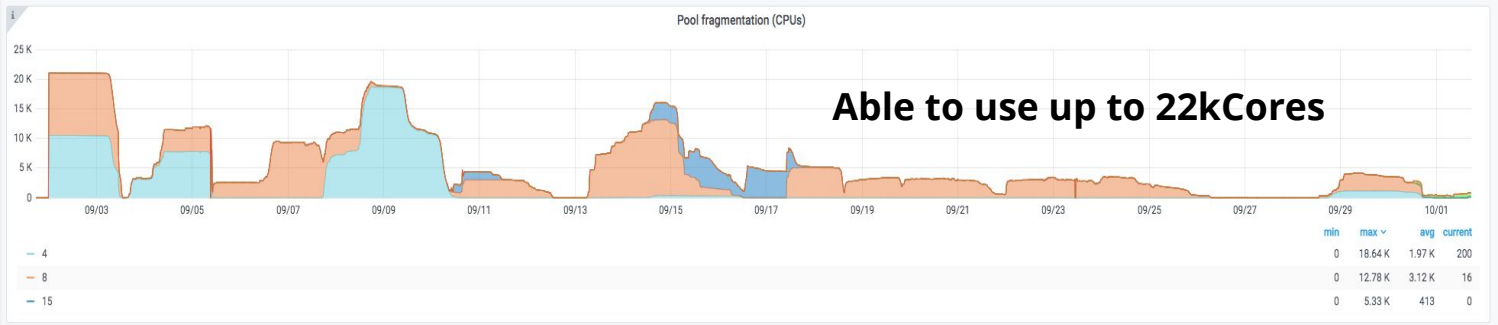
Second goal (Q4 2020 - Q1 2021): large-scale exploitation of CINECA computing resources (mainly through single-core jobs: ~70 jobs per node)



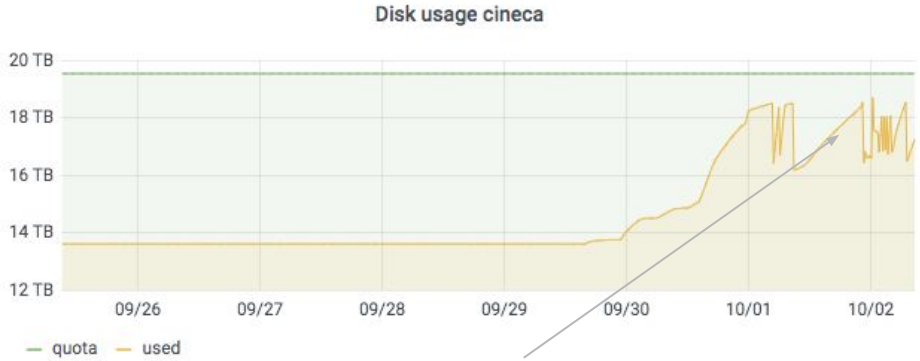


Feb 2021: CINECA was the largest LHCb site outside CERN for detailed MC Simulation

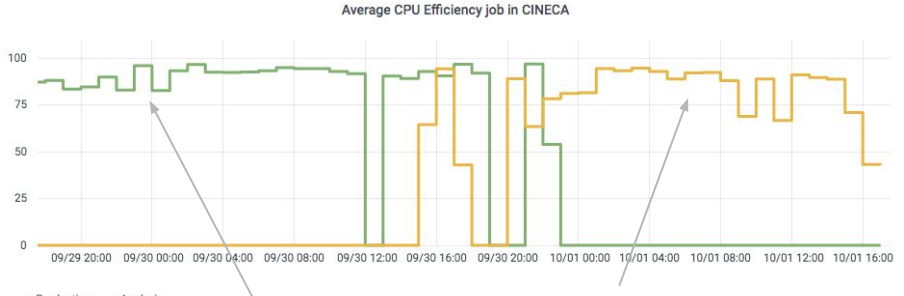
CMS



Able to use up to 22kCores

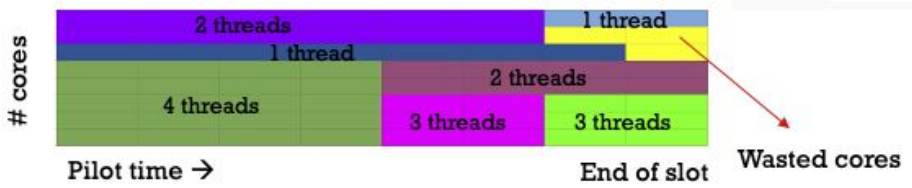


Tested the deployment of a SSD frontend cache (just 20 TB); obviously too small but with a nice test of flushing policies



Able to process central workflows and chaotic user analysis, with very good CPU efficiency (>80% in all the cases).
All CMS workflows can run @ CINECA!

Deployed for the first time in production a local mechanism for cherry-picking jobs in the CMS global pool. CINECA is invisible to the CMS central operations, and appear just as an elastic extension of CNAF



ATLAS @ INFN-CINECA HPC

- Activity started on April 2020
- Montecarlo fast and full simulation jobs (low I/O)
- 48 threads per job, to allow 2GB RAM per thread

Peak usage of 12K simultaneous slots

95% job success rate, including testing periods, very good result

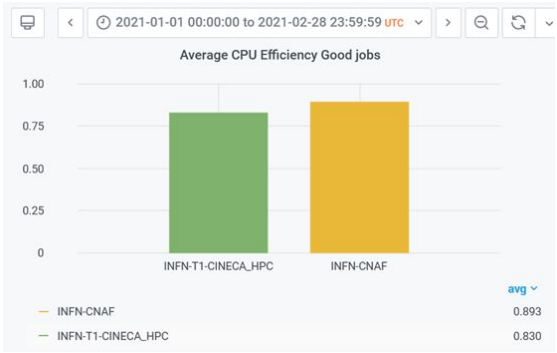
ATLAS @ INFN-CINECA HPC

ATLAS Panda queue created to include CINECA-HPC resources, as part of INFN-T1 site.

- During the grant period, nearly 11% of the computing power used by ATLAS at INFN-T1 was provided by CINECA-HPC resources.

ATLAS standard pilots get input from STORM with file direct access.

- 2020: Pilots modified to get input files with gridftp from CNAF STORM storage .



- 2021: standard pilots used, reading input files with gridftp from CERN EOS storage

MC simulation jobs not IO intensive.

Slight difference in average CPU efficiency comparing with jobs running at INFN-CNAF.

Very small loss in CPU efficiency getting input files from CERN

ALICE … what to run?

* O2 = Online-Offline Project
(ALICE Run-3 software)

PRELIMINARY tests

- Slurm submission → O2* (Run-3) pp simulations
- HTCondor submission → O2 (Run-3) Pb-Pb simulations
- Standard Run-2 submission scheme (pilot jobs, simulation only) to validate the full GRID chain (JDL and register on File Catalogue)

FINAL test

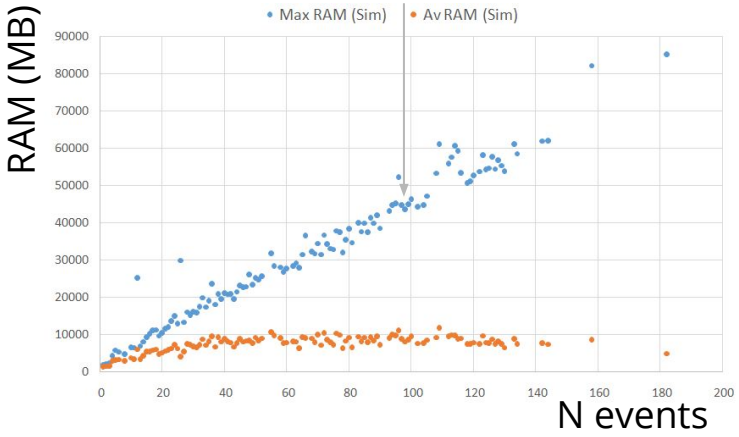
- **O2 (Run-3) simulation to validate the new simulation framework (high parallelization!) on HPC system**

O2 Simulations:

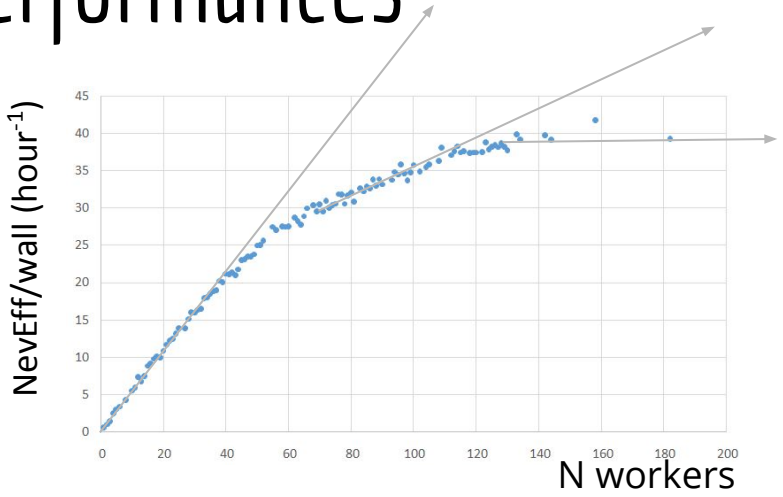
- **100 Pb-Pb events per job (much heavier than pp ones)**
- **with Geant4 propagation: high level of parallelism**
- **From 1 to O(100) concurrent processes (workers)**
- **Digitization (main detectors involved, in particular TPC)**
- **Reconstruction (of simulated data)**
- **Output (~15 GB per 100 Pb-Pb job including everything) written at CERN**

Characterization of performances

N events = N workers (= N procs)
in these simulations



O2 version
23/08/2020



Fine scaling up to the number of physical core (68).
30% gain with HT (up to 120 workers)

$$NevEff = Nev \frac{< event\ size >_{current\ sim}}{< event\ size >_{all\ sims}}$$

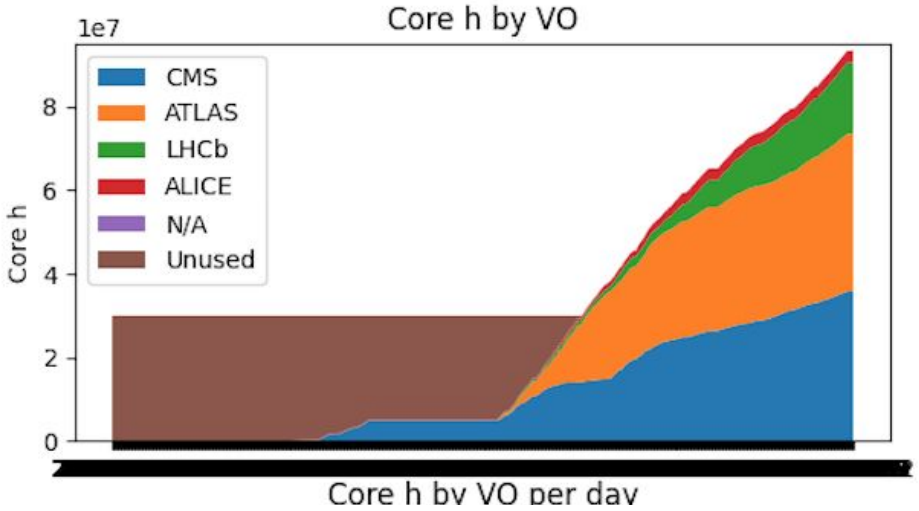
NevEff defined to remove fluctuations due to different seeds used in simulations

Outcomes

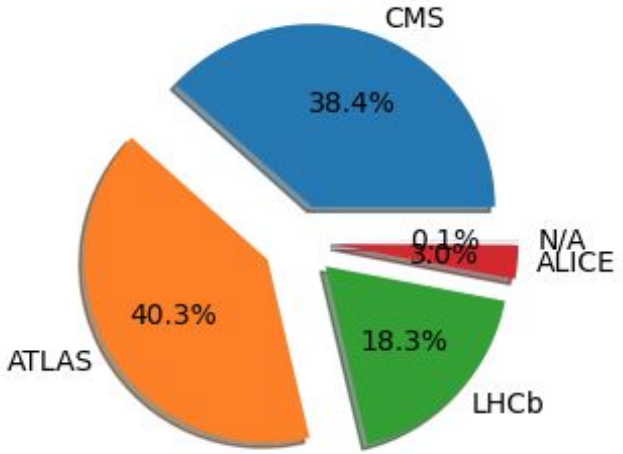
- All the experiments succeeded at using Marconi A2 for the production activities they focused on
- Working with CINECA was very valuable to understand the problems of running LHC workflows in world-class HPC clusters, and it will be helpful to plan and deploy future systems
- In particular:
  - **ATLAS**: successful at adapting MC workflows, further work needed to adapt other types
  - **ALICE**: Validation of the simulation on HPC system; Reconstruction code will run also on GPUs at Run-3; concerns about perf of IO storage access from HPC system
  - **LHCb:** Successful commissioning of new workflows and massive usage of HPC resources, enabled by Dirac
  - **CMS**: massive usage achieved running production jobs; all the workflow types can successfully run.
  - **INFN-CNAF:** gained important insights to plan the layout and capabilities of the new technopole datacenter under construction

The endgame....

- The utilization of the machine was possible beyond the end of the PRACE grant period, thanks t CINECA
- **93 total MCoreH used**




```
Total used is  93459347 hours out of 30M, which is 311 %
It is distributed like:
ATLAS:  125 %
CMS  :  119 %
ALICE:  9 %
LHCb :  56 %
```



How to go on

- By different paths, we already had previous experience on other CINECA clusters
  - **Marconi A3**: a skylake system
  - **Galileo**: CMS small grant in 2020 (used to test network solutions; see M. Mariotti, next talk, https://l.infn.it/bs)
  - **DGX-A100**: ISCRA Grant
  - **Marconi100**
    - 2021: LHC-Italy got 3.5 MCoreH (~ 20 nodes)

MARCONI - 100

Nodes: 980

Processors: 2x16 cores IBM POWER9 AC922 at 3.1 GHz

Accelerators: 4 x NVIDIA Volta V100 GPUs, Nvlink 2.0, 16GB

Cores: 32 cores/node

RAM: 256 GB/node

Peak Performance: ~32 PFlop/s

Quick startup guide

(Tentative) Plans for the experiments on Marconi 100

The italian components of LHC experiments are planning activities for Marconi 100

- **ALICE**
  - Preparing the central code for Power9
- **LHCb**
  - Using the resource for analysis level tasks (with or without the GPUs)
  - Eventually have the central code built for Power9

- **CMS**
  - Code ready for Power9 + GPU
  - System already tested for central analysis jobs
  - Next step: test of central workflows
  - Aiming for physics validation of the platform for CMS
- **Stay tuned for results @ next ISGC**

"Performance" in A.U. of the HLT application (the higher the better)

Thanks to LHCb for lending the node!

In contact with IBM to optimise compilation flags

Platform	CPU only	GPU Type	CPU + 1 GPU	CPU + 2 GPU	CPU + 4 GPU	HS06 CPU
2x Intel Xeon 6130	24	T4	33 (+37%)			865
2x EPYC 7502	58	T4	75 (+29%)	75 (+29%)		1832
2x EPYC 7742	100	T4	127 (+27%)	129 (+29%)		3170
2x POWER9 (Marconi 100 Node @ CINECA)	18	V100	23 (+28%)	23 (+28%)	23 (+28%)	need new compiler flags

CMS High Level Trigger code running on x86 and Power platforms, with GPUs

Conclusions

- By using various opportunities offered to its physicists, from PRACE grants to institutional agreements, INFN has been able to test CINECA system for HEP processing
- The Marconi A2 KNL system was brought into production by the 4 LHC experiments, with considerable success
- The amount of core hours used is visible in all the experiments, but we consider a bigger success the capability to handshake with an HPC site, and to be able to find solutions for all the non trivial technical problems
- The experience we gained will be invaluable for utilization of the Marconi100 system, and later of the Leonardo pre-exascale system

Credits - Apologies: many are surely missing!!!!

LHC-Italy

- Tommaso Boccali
- Alessandro de Salvo
- Concezio Bozzi
- Mirko Mariotti
- Stefano Perazzini
- Francesco Noferini
- Anna Lupato
- Alessandra Doria
- Alessio Gianelle
- Daniele Spiga
- Diego Ciangottini
- Daniele Bonacorsi
- (...)

CNAF

- Stefano Dal Pra
- Stefano Zani
- Gaetano Maron
- Luca dell'Agnello
- Lucia Morganti
- Daniele Cesini
- Vladimir Sapunenko
- (...)

CINECA

- Massimiliano Guarrasi
- Marcello Morgotti
- Daniela Galetti
- Carlo Cavazzoni
- (...)

CERN

- Andrea Valassi
- Federico Stagni
- (...)