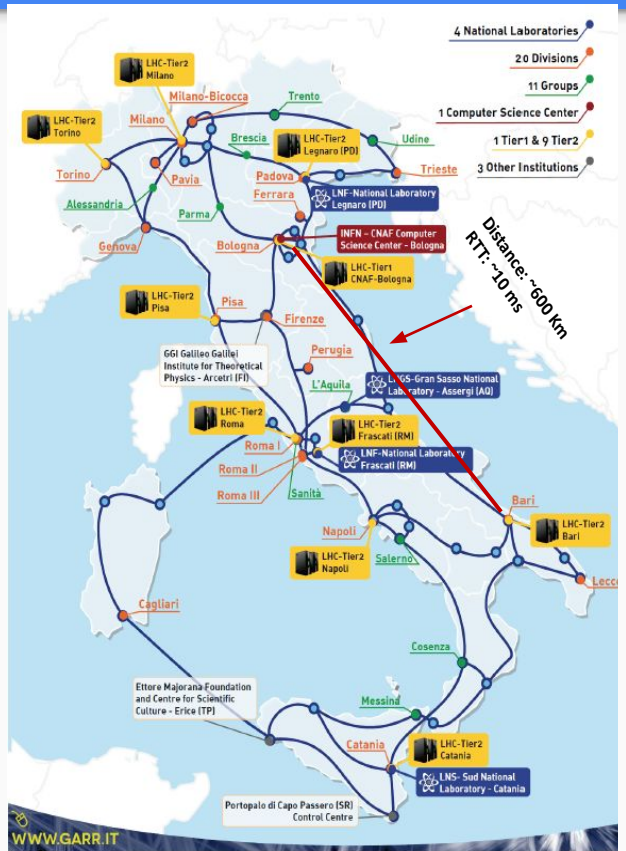
Experiences in integration with HPC @ INFN

Daniele Cesini (INFN-CNAF)
Giacinto Donvito (INFN-Bari)
Cristina Duma (INFN-CNAF)
Daniele Spiga (INFN-Perugia)
Tommaso Boccali (INFN-Pisa)

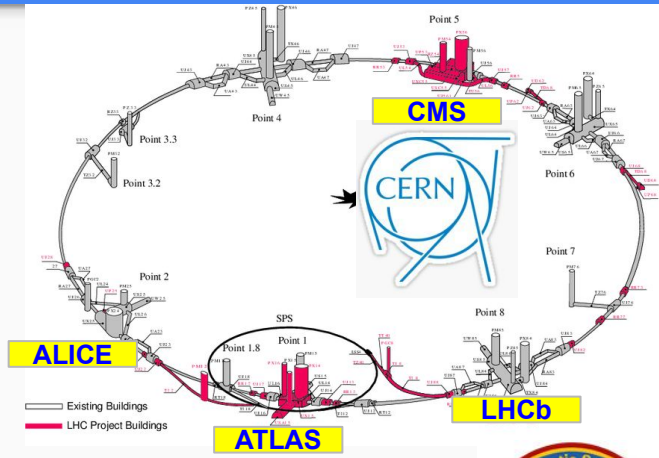
INFN computing at a glance

- INFN has a long history in Scientific Computing, starting from the first clusters, the first attempts at multi-site computing, the realization of home-made HPC with APE, to the current distributed structure, using several “owned” centers
  - CNAF (Bologna), which hosts the INFN Tier1 for WLCG
  - 9 sites host the WLCG Tier2s
  - PON funded centers (ReCaS, IBiSCo)
  - Other smaller sites host departmental farms (or Tier3s)
- The Tier1 accounts for nearly half of total computing and storage resources dedicated to INFN experiments
- Computing for theoretical physics is today mostly performed on HPC resources at CINECA (the Italian EuroHPC Prace Tier-0 node), which is just 8 km away from CNAF



Which computing?

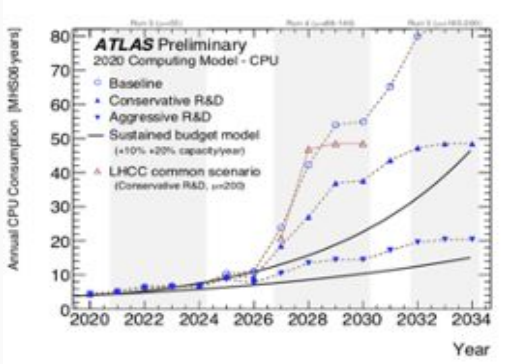
- As quite typical for institutions supporting High Energy Physics, the largest resource share (some 80%) is used to support LHC experiments, with a ~10% share of WLCG worldwide resources (~100,000 cores, ~100 PB disk, ~100 PB tape)
- This “bulk” of utilization is pretty standard: x86_64 CPUs, no multinode payloads, submission via batch systems



- On top of this, more modern / diverse use cases are starting to appear, including the needs to access accelerators, parallel processing, interactive environments
- **This is where HPC systems become interesting, and the main reason for a number of INFN experimentations on HPC systems**
- On top of this, some HPC systems are not-so-different from our standard machines, and there can be economic interest in using them

The computing problem ahead of us

- High Energy Physics (INFN largest commitment, via WLCG) has been in the last ~20 years the scientific domain with largest computing needs. Somehow we survived but we had to learn, develop and deploy solutions which were not in our portfolio:
  - Distributed computing
  - The GRID
  - The Cloud
- In the next decade we expect HEP to be still dominant, with some frightening extrapolation -- but not alone: astroparticle experiments (Virgo/Ligo, SKA, CTA, …) are going to be at least on par, with a larger use of technologies like GPUs. We need to get ready now



These points up here means "we need more money" (*forget it*)
This band is the "constant money region" (*we can afford it*)

HPCs are seen as (one of the) (possible) solution to both problems

a. **If we are able to translate Flops in HS06**
b. **If we are able to do data intensive and distribute processing at HPCs**

ProtoDune 2-3 GB/s (like CMS); Real Dune 80x

SKA up to 2 PB/day

A single genome ~ 100 GB. a 1M survey = 100 PB

CTA projects to 10 PB/y

What does INFN have access to?

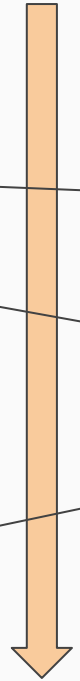
- **Owned resources @ CNAF: two local clusters**
  - O(1k) HT cores + 10 NVIDIA GPUs - infiniband interconnect
  - O(1k) HT cores - Omnipath interconnect
  - Dedicated storage on GPFS

- **Resources at or close to our sites, coming from national / regional projects:**
  - PON iBiSCo - Bari (about 20kCores + 50 NVIDIA), accessible via HTCondor or Docker orchestrators

- **Other Resources we can access as "users":**
  - CINECA EuroHPC/PRACE Tier-0 center is just 8 km from CNAF
  - Since 2018, we have deployed a dark fiber between the sites, operated via an infinera CloudXpress DCI (up to 1.2 Tbps) → CINECA resources are "same as local" for CNAF
- In the near future (2022), INFN is one of stakeholders of the italian Pre-Exascale Leonardo @ CINECA

… and what can we do with them?

1. HPCs as an R&D platform for complex experiment workflows needing parallel computing and accelerators
2. HPCs becoming provisionable via Cloud Interfaces, as the other types of resources
3. HPCs as (on demand) backends to batch / interactive computations
4. HPCs as a part of the “bulk” of experiment processing, with the site-extension mechanism; the accelerated (if present) part can be used “somehow”
   a. By GPU enabled workflows
   b. Using job composition on the same slot (1 CPU intensive and 1 GPU intensive workflows paired)



From “easy” to “hard”, but also from “low gain” to “high gain”

Use HPC as “generic resources”

Can open new possibilities for our users in an era where ML and Jupyter are becoming dominant for analyses

Can in principle adsorb the bulk of INFN computing needs, for internal activities and international commitments

(focussing on the last 3 aspects in the following…)

HPCs as part of the standard distributed computing

- Production level activities demonstrated with the PRACE Grant #2018194658 (PI: T.Boccali) on the **Marconi A2** (KNL) and **Galileo** (Xeon) HPC systems
- Platforms are x86_64, so no real problem with sw, but a lot of work to overcome the mismatch between typical WLCG jobs and a typical HPC site

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**

GALILEO

UserGuide

Model: IBM NeXtScale cluster
Architecture: Linux Infiniband cluster
Network: Intel OmniPath (100Gb/s) high-performance network

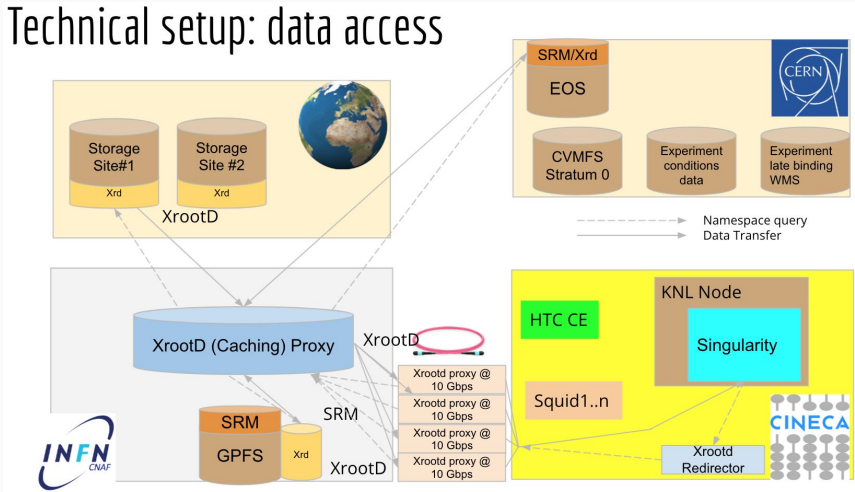
Nodes: 1022 (Intel Broadwell)
Processors: 2 x 18-cores Intel Xeon E5-2697 v4 at 2.30 GHz
Cores: 36 cores/node, 26.572 cores in total
RAM: 128 GB/node

Deploy edge services at the boundary CINECA/CNAF: HTCondorCEs for SLURM, SQUIDs for Database access, ...

A standard CINECA Marconi A2 is node configured with	A typical WLCG node has
An Intel(R) Xeon Phi(TM) CPU 7250 @ 1.40GHz: 68 or 272(HT4x) cores, x86_64, rated at ~¼ the HS06 of a typical Xeon per core	1-2 Xeon-level x86_64 CPUs: typically 32-128 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 (so a total 64-256 GB)
No outgoing connectivity from the node	Full outgoing external connectivity, with sw accessed via CVMFS mounts; additional experiment specific access needed (condition DBs, input files via remote Xrootd, ...)
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 (via passport / fiscal code identification)	Access via pilots and late binding; VOMS AAI for end-user access

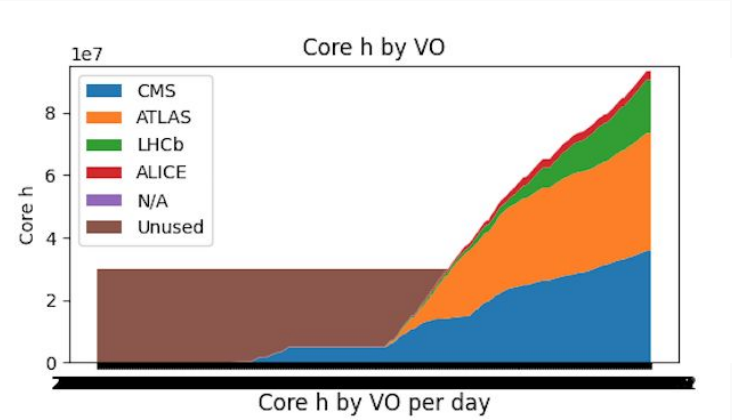
Enable NATted connectivity to just CNAF and CERN, use proxy-caches to fan-out

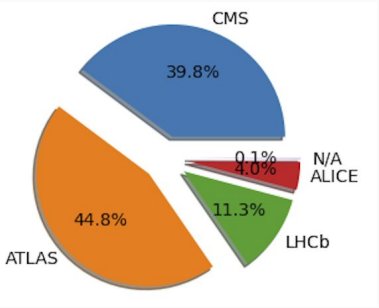
Technical setup: data access



Marconi A2 usable for production at WLCG

- With the modifications in the previous page (and more) we were able to bring into production workflows for the 4 LHC experiments (here)
- 93McoreHours used by the experiments from the grant





- **Major goals reached:**
  - HPC centers are _not_ incompatible with the data intensive / distributed processing of High Energy Physics, with nearly optimal performance (not shown here)
  - HPC centers can be used as a dynamic extension of an existing site (CNAF in this case) without additional load on central experiment operation team
  - Record is 22kCores used at one moment by CMS … which means the “CNAF-extended” was ~ 4x standard CNAF

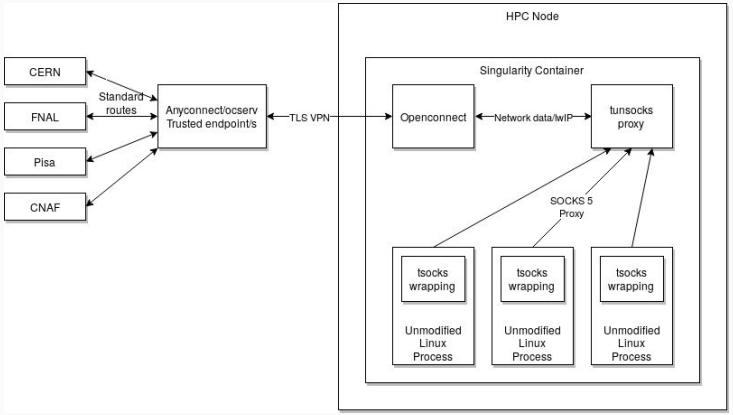
With some unexpected future oriented R&D →

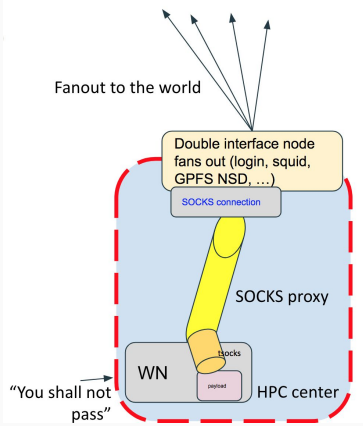
… a possible portfolio of tools to interface HPCs

- A “side effect” of the production level utilization of CINECA HPCs by INFN has been the preparation of a **portfolio of solutions** which are not strictly specific for the effort:
- **Proxying-Caching layers** used not only to increase CPU efficiency, but also to fan-out connections to/from sites normally not accessible
- **Edge services** (Squid, HTCondorCE) to be used as a bridge between the two infrastructures; eventually deployed as containers
- And an interesting **low level tool for networking**: with Marconi A2, we were able to handshake with CINECA sysadmins connectivity to CERN/CNAF, **but what to do where it is not the case?**

- We started looking for a tunnelling solution which:
  - … can be deployed (without hacking the system) by a standard user
  - … covers all the possible connections (UDP, TCP) and services (XrootD, SRM, HTTP(S), …)
  - … is not intrusive for SW (no recompilation, no changes of configuration, no need for special workflows)
  - … does not require Linux namespaces to be allowed (they are not available @ CINECA)
  - If you want, an universal edge service working below the application layer
- After some research, 2 possible solutions were found:

1. tsocks + ssh + cvmfsexec + (singularity)
2. tsocks + tunsocks + openconnect + cvmfsexec + (singularity)



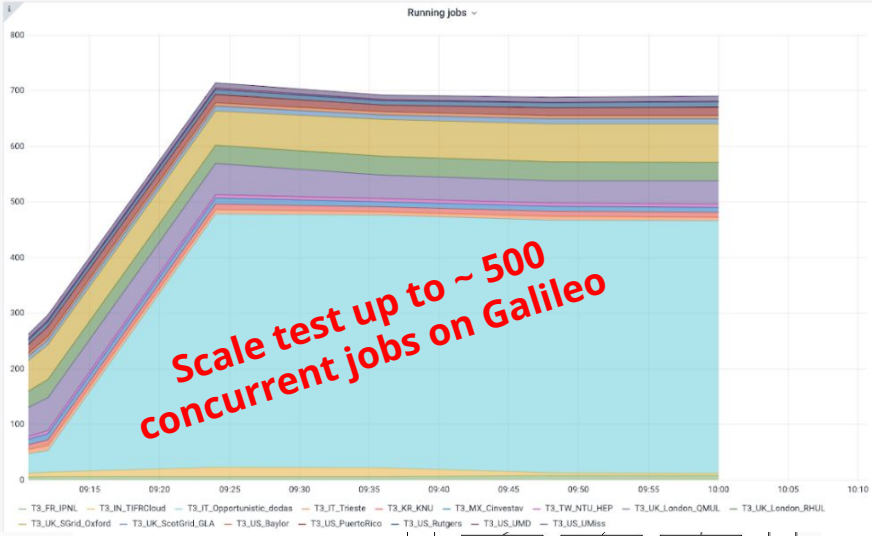


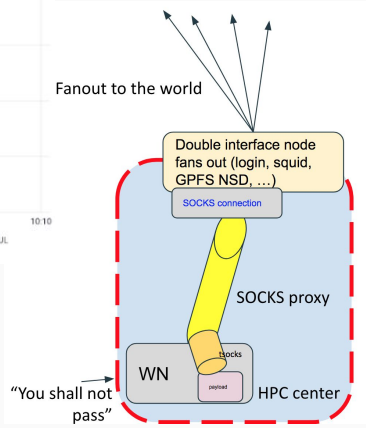
Tested at scale on CINECA’s Galileo (with no external connectivity from the computing nodes)

… a possible portfolio of tools to interface HPCs

- A “side effect” of the production level utilization of CINECA HPCs by INFN has been the preparation of a **portfolio of solutions** which are not strictly specific for the effort:
- **Proxying-Caching layers** used not only to increase CPU efficiency, but also to fan-out connections to/from sites normally not accessible
- **Edge services** (Squid, HTCondorCE) to be used as a bridge between the two infrastructures; eventually deployed as containers
- And an interesting **low level tool for networking**: with Marconi A2 we were able to handshake with CINECA sysadmins connectivity to CERN/CNAF, **but what to do where it is not the case?**

- We started looking for a tunnelling solution which:
  - … can be deployed (without hacking the system) by a standard user
  - …nd services (XrootD, SRM, HTTP(S), …)
  - …nges of configuration, no need for special
  - …d (they are not available @ CINECA)
  - …w the application layer
  - …nd:
  - …larity)





Tested at scale on CINECA’s Galileo (with no external connectivity from the computing nodes)

Getting more complicated: Marconi 100!

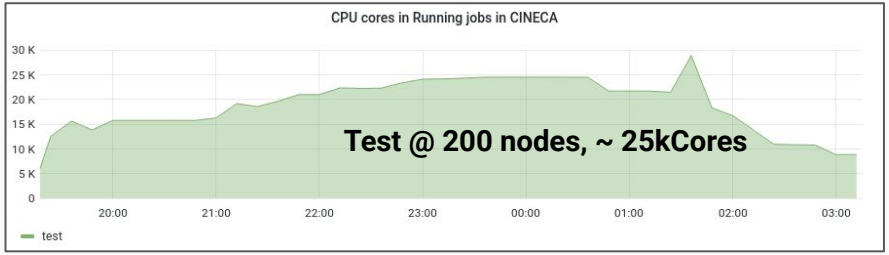
MARCONI - 100

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

- CINECA Marconi 100 (M100) is far more complex to use (and hence more interesting!)
  - **32(x4) Power9 cores per node - cannot run x86_64 code**
  - **4 Nvidia V100 per node - performance comes from GPUs**!
- So you need at least
  - To have **Workload Management Systems** which are architecture aware
  - To have SW compiled for ppc64le (or, architecture agnostic as Python can be)
  - To have a way to use the GPUs
    - Either via production workflows
    - Or as paired workflows to standard CPU only production workflows

CMS@CERN was in the best position for the test:

- SW is already compiled for ppc64le (although not validated)
- We could easily modify the WMS to be architecture-aware
- The WMS client is Python only
- CMS has “some” GPU friendly workflows at least for tests - but not enough!



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

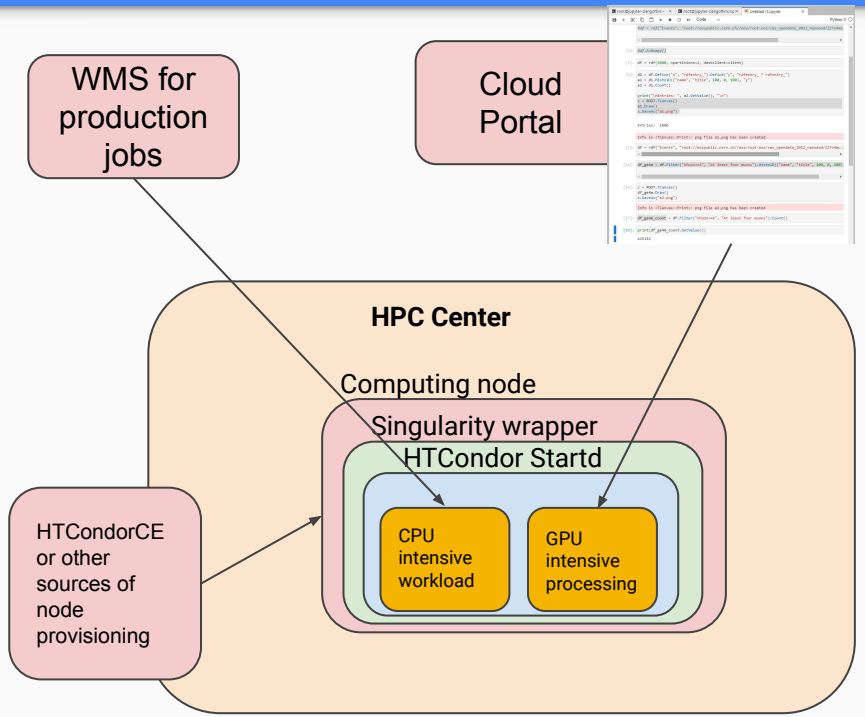
Thanks to LHCb for lending the node!

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

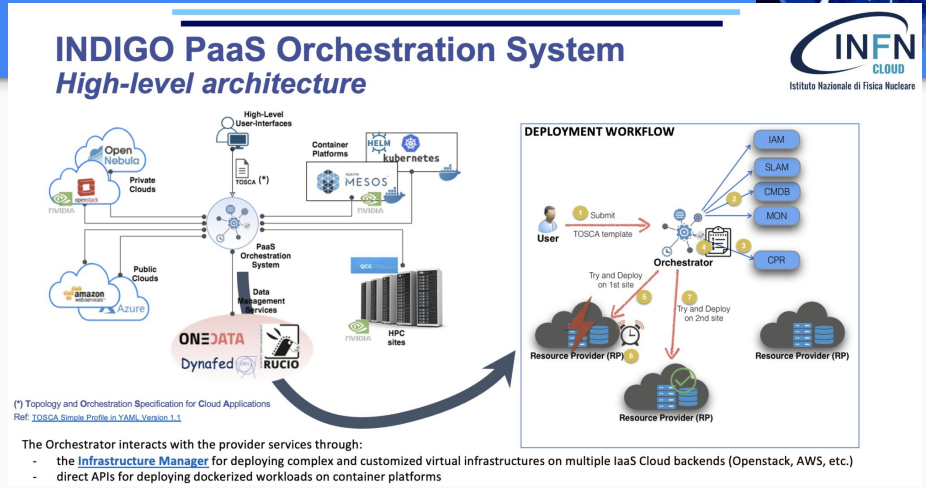
What to do with GPUs?

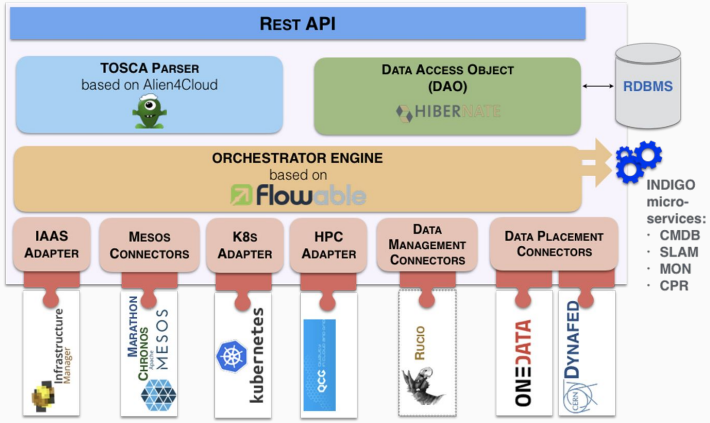
- M100 and Leonardo have tons of GPUs. Too many for our standard workflows
- On the other hand more and more INFN users are searching for interactive or batch GPU resources for interactive analysis (Jupyter Notebooks, Numpy, Pandas, Scipy, Sk-learn, Keras+Tensorflow, PyTorch, …)
- Idea is to match the use cases:
  - Schedule a “mostly CPU standard workflow” and either a Jupyter Notebook backend or a Python ML training (for example)
  - Potentially provide every INFN user with a perfect environment for data analysis: an HPC class node with 1-4 top class GPUs



HPC access via Cloud orchestrator

- The PaaS Orchestrator supports the deployment of virtual machines and containers that need to access specialised hardware devices, namely GPUs, to provide the processing power required by tasks like Machine Learning algorithms
  - the GPU requirements (num, vendor, model) can be specified in the TOSCA template
  - the Orchestrator automatically selects the sites/services that provide the needed capabilities (flavors, GPU support)
- The Orchestrator includes a plugin for submitting jobs to HPC facilities
  - exploits the QCG-Computing service (PSNC) that exposes REST APIs to submit jobs to the underlying batch systems





Conclusions

- INFN computing is in continuous evolution, and it crosses paths with HPCs more and more frequently - via internal and external resources
- In particular, the geographical and political vicinity with CINECA and the common handling of the Leonardo Pre-Exascale systems is a peculiar situation which will drive future use cases
- We are working towards a as-close-as-possible integration of HPCs in our systems, including bulk processing, cloud instatiantion, interactive utilization
- While doing so, we have prepared tools / solutions / documentation which constitute a sort of toolset we are happy to share (*as already happening in some cases…*)
- Our next adventure: