Uso di HPC negli Esperimenti LHC

Daniele Spiga INFN-PG
Tommaso Boccali INFN-PI
On behalf of

Stefano Piano, Francesco Noferini, Concezio Bozzi, Leonardo Carminati, Alessandra Doria, Alessandro De Salvo, Lorenzo Rinaldi

Outline

Why we talk about HPCs integration @HEP

What are the Integration challenges

LHC Experiment: status and ongoing activity

- A quick look at national landscape

Summary

The HEP (and close friends) Computing circa 2020

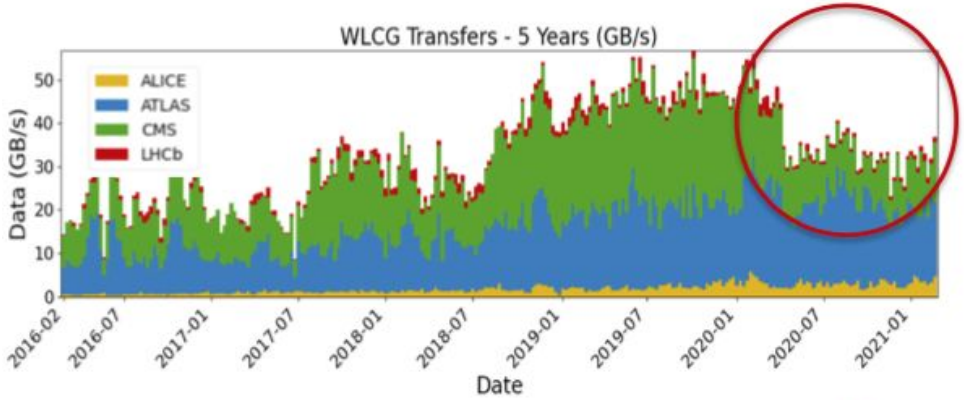
- Since ~1980, the HEP world has been facing a steady increase in the computing needs, at least from LEP times
- The increase in needs has seeded most of INFN Computing R&D and operations in the last 20 years
  - The GRID
  - The Tier-1 and Tier-2 centers
  - Now the Cloud, the Datalake, ...

	ALEPH 1995 	CDF 2004 	CMS 2007 
Dimensione dei dati raccolti	**1 TB** = 1000 GB	**1 PB** = 1000 TB x1000	**~10 PB** x10
Capacita' di calcolo (SI2k)	**<<100k**	**1.4 M** x50	**>25 M** x20

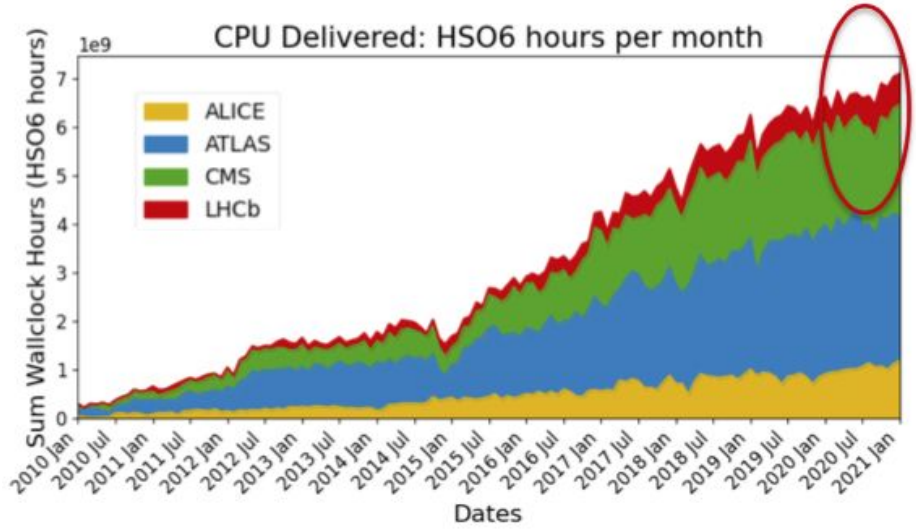
Nota: 1 PC attuale ~ 1 SI2k

2021: WLCG

- 161 official sites, in 42 countries
  - “We only miss Antarctica”
- Pledged resources
  - ~ 8 MHS06 (~800kCores)
  - ~700 PB disk
  - ~1 EB tape
- Other resources (HLT farms, overpledges) count for at least another 50% on CPUs
- Transfers (as seen by FTS) ~ 50 GB/s



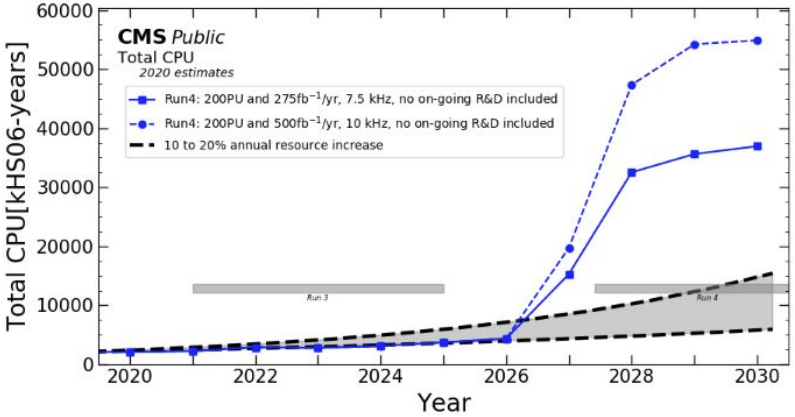
LHC is not running now

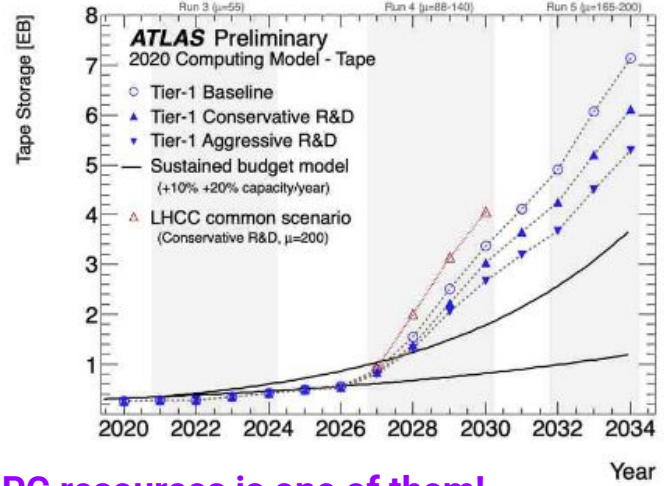


7e9 HS06h ~ 800kCores 24x7

What is in front of us?

- LHCb + ALICE upgrade already happened, start data taking in ~ 10 months
  - Ambitious, but manageable in semi-adiabatic mode
- ATLAS and CMS ~ 2028 with Phase-2
  - Currently still unable to match a flat funding profile

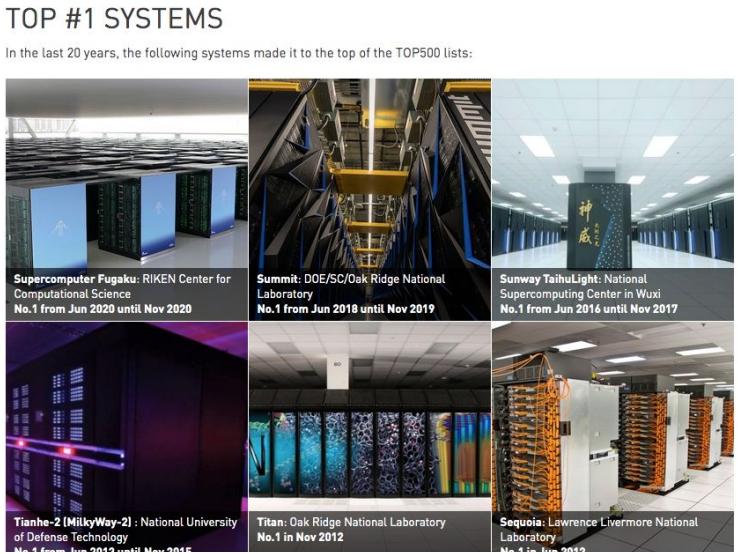


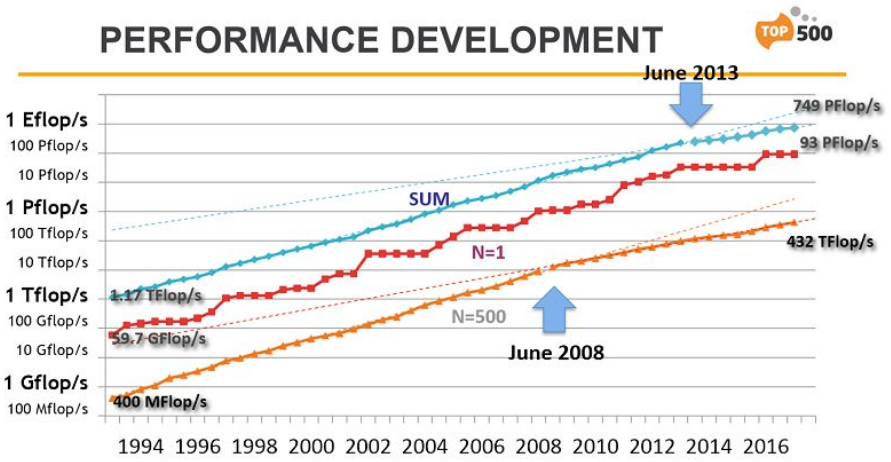


Many solutions / ideas under test; using HPC resources is one of them!

The HPC (“Supercomputer”) panorama

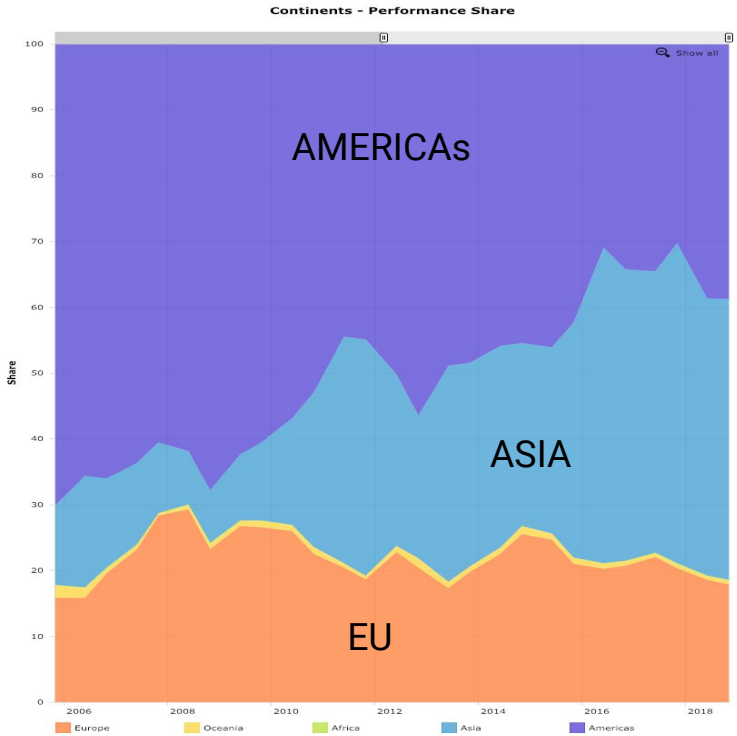
- The world is literally full of Supercomputers. Why ?
  - Real scientific use cases
    - Lattice QCD, Meteo, ...
  - Industrial showcase
    - And hence not 100% utilized, opportunities for smart users. Can we be one of them?
- They are usually very closed systems
  - Few users, with systems designed for their benefit
  - Highly parallel: the cost of networking is a big part of the total
- They are expensive
  - 100M-1B$ each is the typical range (for comparison, the total “cost” of the WLCG resources is ~ 100 M$)
- Some hint of global slowing down, but not for top systems where the “war” is on





Supercomputer - The expected future

- The race will go on, at least between major players
- EU wants to enter the game - never at the top in the last 25y
- Next big thing is **ExaScale** (10^{18} Flops - operations per second)
  - Should be well available by HL-LHC times; indeed in principle by this year!
- Somehow difficult to compare, technologies / benchmarks, but
  - LHC needs today the equivalent of ~30 PFlops
    - Not all the flops are equal! This is CPU currently
  - A single Exascale system is ok to process 30 “today” LHC
  - **Scaling: a single Exascale system could process the whole HL-LHC with no R&D or model change**
- Some FAs/countries are explicitly requesting HEP to use the HPC infrastructure as ~ only funding; **it is generally ok IF we are allowed to be part in the planning (to make sure they are usable for us)**



Top machines now

- 7.3 MCores, with ARM architectures
- 5 PB of RAM
- RedHat Linux
- 29 MW (~60 M$/y ??)

The second system is

- 2.4 M computing cores - Power9 + NVidia V100
- 2.8 PB ram
- RedHat Linux 7
- 10 MW power use (which alone is ~ 20 M$/y)

It is difficult to translate this power in the HS06 metric we use, but:

- **Intel Xeon Xeon E5-2640v3 is rated at 350 GFlops**
- **The same CPU is rated at 230 HS06**
- **→ 1 HS06 ~ 1.5 GFlops**
- **Fugaku ~ 250 MHS06 (which is ~ 40-60x WLCG)**

A quick look at HPCs main features

HPC design features:

- Performant node-to-node interconnection
- Strict user access policy (used == ID card)
- Scarce / absent local scratch disk
- Limited capability for data access outside the facility, if not absent
- Relying on accelerator to boost performance (today mostly Nvidia GPUs, tomorrow FPGA, Intel Xe, AMD, … ??)
- Storage systems are optimized for latency and speed, and not for total size.

Moreover

- HPC centers differ on a variety of specialized hardware setups and policy wise strict usage rules can apply mostly for security reasons

All experiments are working on integration of HPC resources with varying levels of technical difficulties because the HEP systems are built using different technical needs in mind:

- HEP workflows are typically data intensive, and systems deploy large storage systems close to the computing, balanced with CPU deployment. Most of the software is still designed and optimized for the x86_64 architecture

Let's see a "good" scenario [an example]

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

A detailed analysis by category

(a longer list..)

https://cds.cern.ch/record/2707936/files/NOTE2020_002.pdf

Category	Explanation	CMS standard solution	CMS preferred solution for HPC	CMS fallback workable solution (full utilizability)	CMS fallback solution (for a fraction of workflows)	CMS no-go scenario	Possible CMS devels to solve the no-go
Architecture	Base system architecture	x86_64	x86_64	x86_64 + accelerators (with partial utilization)		Currently, OpenPower, ARM, … they could be used but at the price of physics validation	QEMU? Recompiling + physics validation?
Memory per Thread/core	Memory available to each thread / process	2 GB/Thread	2 GB/Thread	Down to 0.5 GB/thread needs heavy multithreading, at the expenses of CPU efficiency	GEN and SIM workflows need less than 2 GB/Thread (0.5GB/Thread would be a limit) in order to run efficiently	Less than 0.5 GB/thread	
I/O	I/O demand per process	5 MB/s/core	5MB/s/core		GEN and SIM workflows are mostly CPU bound still ok with 0.1 MB/s/core	Less than 0.1 MB/s/core	
Local Scratch space	Local space per production job	20 GB/Thread	20 GB/thread local	Less than 20 GB/thread ok if a shared high performance FS is available on all the machines Large multithreading lowers 20 GB/thread requirement to ~ 10	Some CMS workflows run for hours without creating huge local disk areas (GEN, SIM)	No sizeable local space and no shared usable FS	
Outgoing networking	Needed on WNs in order to access remote data, conditions, and to speak to the CMS Global Pool	Full outgoing connectivity	Full outgoing connectivity	Connectivity to only a subset of the IP ranges (for example, to CERN, and to a close xrootd proxy cache) And to everywhere we have condor services?	NAT with a very limited bandwidth via an edge service	No outgoing connectivity from the compute nodes and no NAT available	Edge service running Harvester or HTCondor? Prepare a single container to be deployed at the edge and doing: NAT for Condor, Squid, Xroot proxy cache, …?

HPCs integration challenges in summary

Integrating HEP experiment workloads on HPC systems poses technical issues and challenges in two distinct areas:

- Data management and submission of jobs
  - **how to schedule jobs on HPCs, how to connect them with experiments services, how to access experiment software and calibrations ..**

- Development of the software applications
  - **HEP software has so far almost exclusively been developed for traditional x86 CPUs, while HPC systems provide (and will provide more and more) their computing capacity using Accelerators (GPUs /FPGAs) and a variety of technologies such as PowerPC processors**

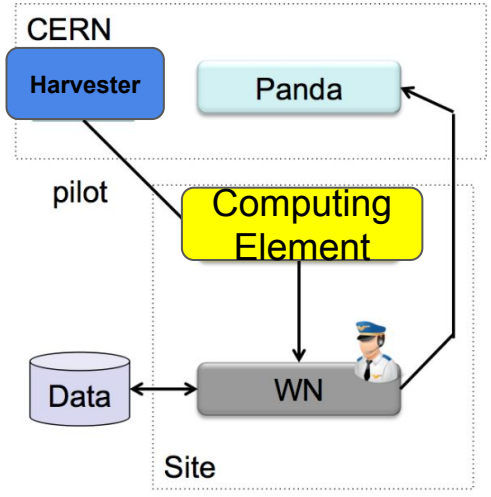
https://zenodo.org/record/3647548#.YKzxqqIzZH5

https://cds.cern.ch/record/2016011/files/actchep2015-linennumbers.pdf

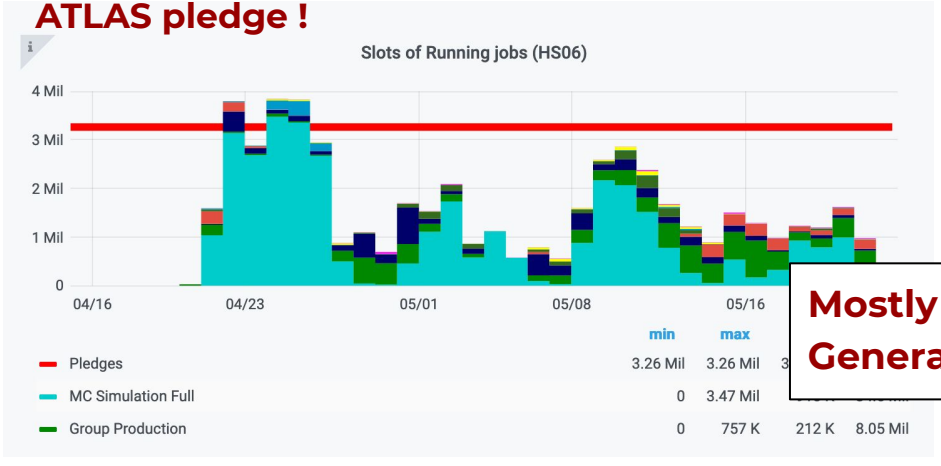
ATLAS: Grid Like execution @HPC

HPC provides the external connectivity on the nodes and enables access to cvmfs software area

- Any ATLAS workflow can be executed. Push and pull mode supported although push mode is preferred on HPCs without close Storage Element. Dedicated agents (ARC-CE or Harvester) are used
- The data on the grid is typically accessed directly with xrootd protocol or copied to/from local node scratch space.



VEGA alone could be able to cover the full ATLAS pledge !



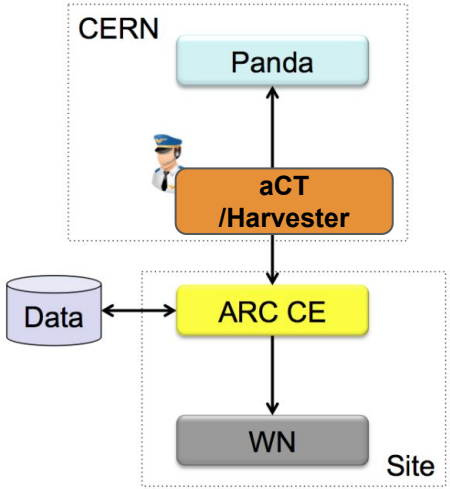
CINECA_HPC covered up to 10% of the total CNAF pledge

Mostly Event Generation

ATLAS: Workflows for ‘closed’ HPCs (the hardest case)

Limited connectivity execution: when the nodes do not have outbound network access.

- The ATLAS software can either be installed locally on a shared file system or provided through fat containers on HPCs that support singularity or shifter. Such HPCs typically run **ATLAS Event Generation or Geant 4 Simulation** where conditions data can be stored in a local SQLite file.
- The ATLAS agent (ARC-CE or Harvester resources broker) receives the payload description including a list of input files. This list is then processed by the agent’s data-transfer subsystem. **Dedicated data-transfer nodes can be transparently used and tuned for transfer performance**. The completed payloads are processed by the ATLAS agent and the output files are then transferred to the pre-assigned remote Storage Elements.



Harvester has been interfaced to multiple Cloud resources (Google Compute Engine, OpenStack), several US DOE High ranking HPCs (Titan, Theta, Cori, US NSF HPC Frontera)

HPC and cloud resources

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

- Thanks to the new O2 simulation and reconstruction code (Run 3) possible to fully exploit the multi process features
- Significant progress has been made to incorporate HPC and cloud resources in the standard ALICE Grid workflows
  - Multicore queues at CERN used to test and benchmark the O2 MC code
  - Intel based HPCs (Marconi @ CINECA, Cori and Lawrencium @ LBNL) have been used for the O2 MC challenge

- Future steps:
  - porting the O2 code to Power 9 and ARM platform

LHCb: distributed computing integration

- **Support HPC with external connectivity** (multi-core allocations)
- Tools to exploit HPC with **no external connectivity are in progress**

Tackling the distributed computing challenges

Overview

- 1 main variable directly affects the chosen solution (push, pull):
  + Do WNs have an external connectivity? yes (or only via the edge node), no
- Other variables generate variations that can be added up to the proposed solution:
  + Is CVMFS mounted on the WNs? yes, no
  + Is LRMS accessible from outside? yes, no
  + What type of allocations can we make? single-core, multi-core, multi-node

⇒ We will go through different cases: from the easiest to the hardest one

Tackling the distributed computing challenges

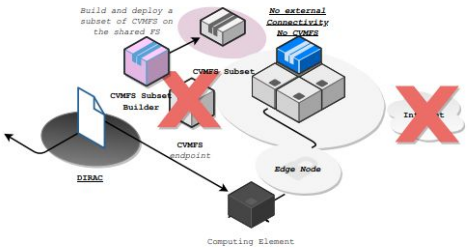
Push model: No Ext. connectivity, No CVMFS

In Progress: v7r2? VO action

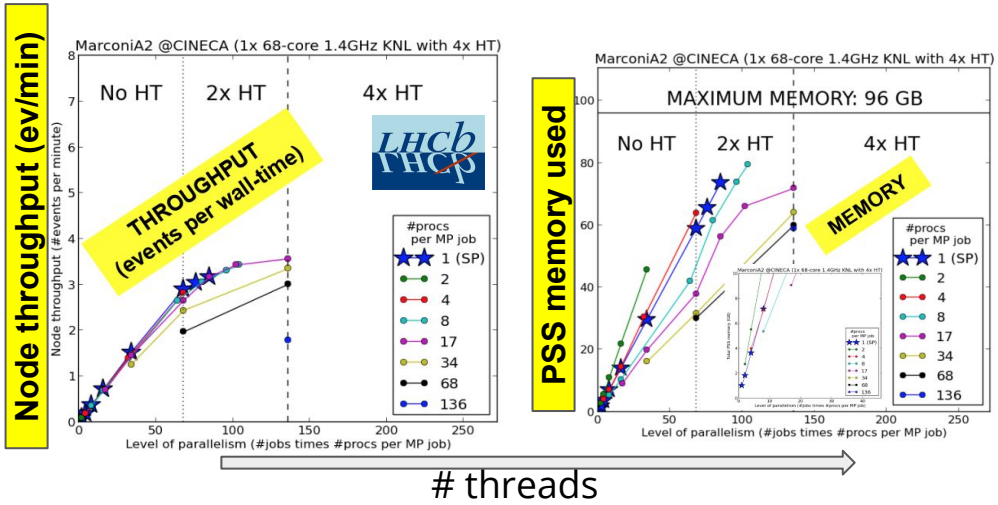
As it was already said, SC do not provide CVMFS by default. CVMFS-exec cannot be used in this context.

Subset-CVMFS-Builder

- Run & extract CVMFS dependencies of given jobs
- Use CVMFS-Shrinkwrapper [1] to make a subset of CVMFS
- Test it & deploy it on the SC shared FS



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

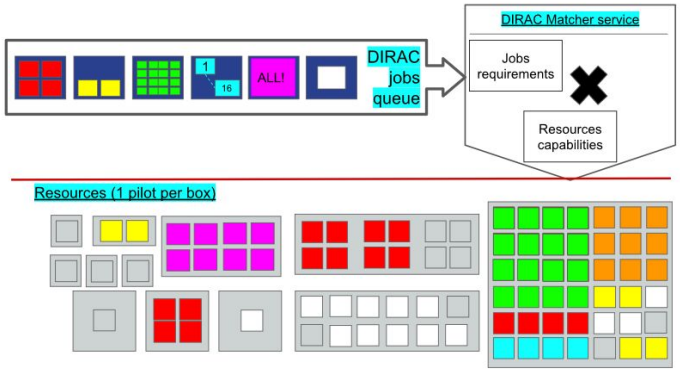


LHCb MC Simulation software

- 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



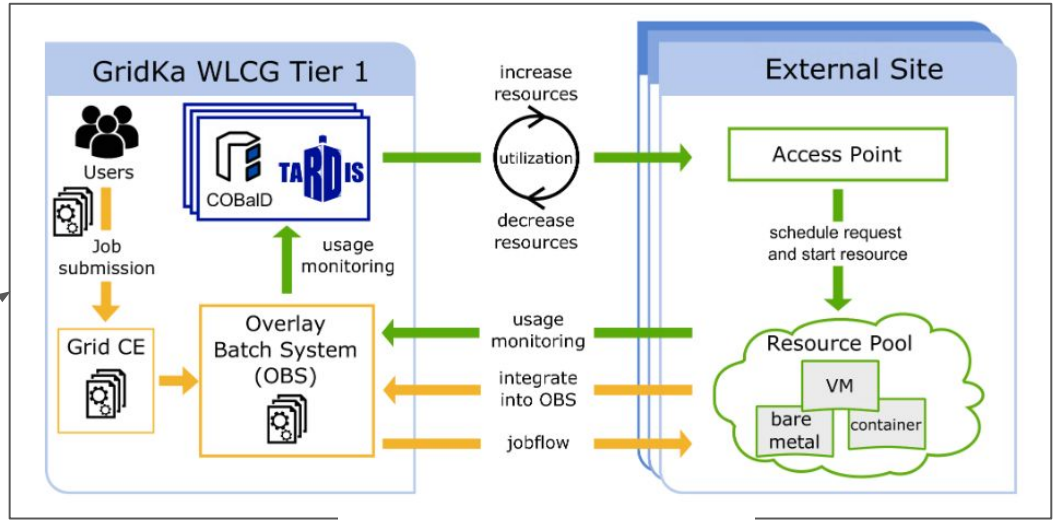
CMS current approaches

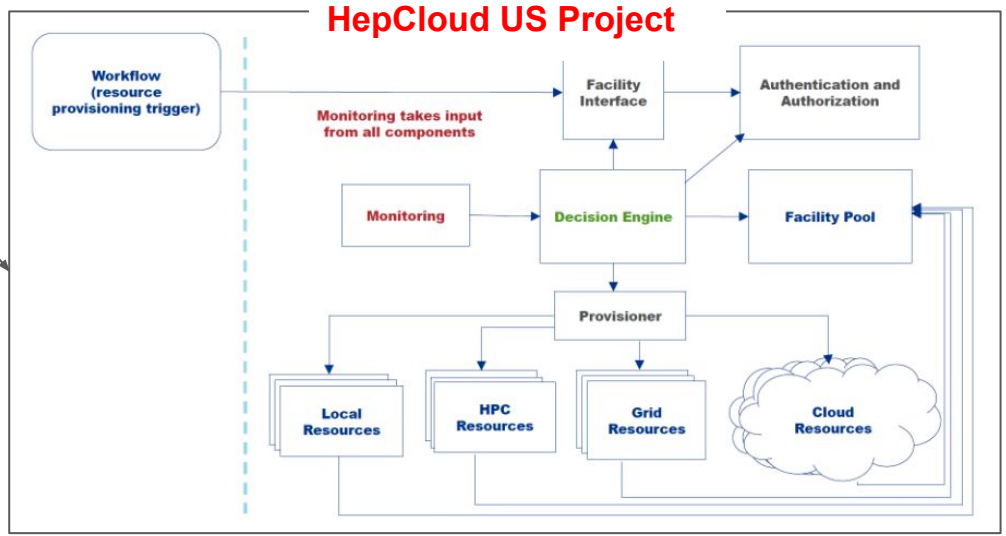
CMS has developed and deployed two distinct strategies

1. **Overlay batch model**
   a. Currently two incarnations

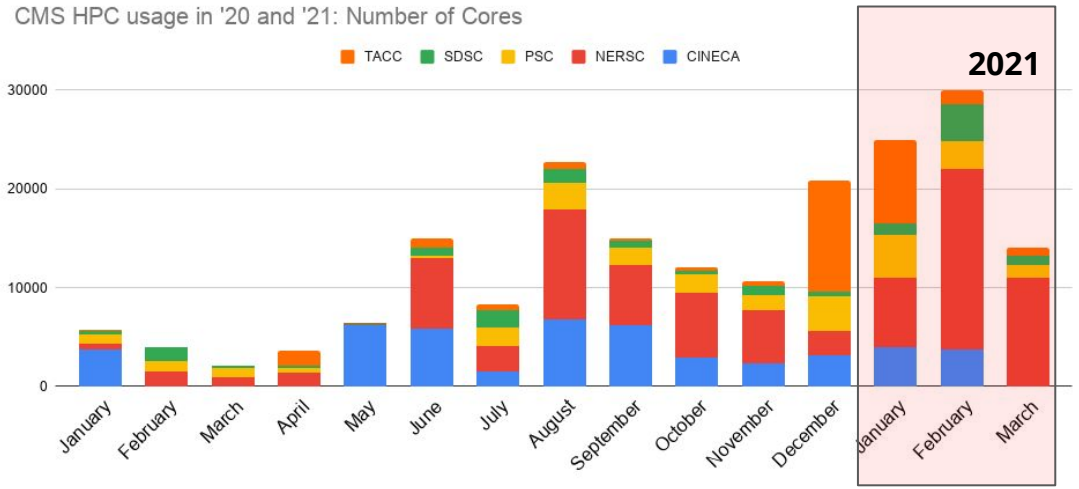
2. **Site Extension**
   a. adopted at CNAF
   b. Mainly relying on two pillars
      i. Storage access
      ii. Site level defined matchmaking rules to allows to cherry peak only suitable workflow.

3. Investigating the HTCondor split starter mechanisms for filesystem based communication (BSC)





HPC Overall usage



Resource providing beyond-pledge capacity	Type	Used CPU power [kHS06]	% of total opportunistic	% of total used compute power
HLT	Cloud	303.0	26.8	9.5
NERSC (HEPCloud)	HPC	19.5	1.7	0.6
PSC (HEPCloud)		27.8	2.5	0.9
SDSC (HEPCloud)		14.3	1.3	0.4
TACC Frontera (HEPCloud)		17.6	1.6	0.5
TACC Stampede 2 (HEPCloud)		3.2	0.3	0.1
TAAC Jetstream (HEPCloud)		12.1	1.1	0.4
CINECA		9.4	0.8	0.3
ForHLR2 (KIT)		1.2	0.1	0.0
CLAIX (RWTH Aachen)		2.7	0.2	0.1
Tier-0 (beyond pledge)	Grid sites	56.3	5.0	1.8
Tier-1s (beyond pledge)		62.4	5.5	1.9
Tier-2s (beyond pledge)		546.1	48.2	17.0
Tier-3s		43.2	3.8	1.3
OSG	Grid	0.4	0.0	0.0
BEER	/	8.7	0.8	0.3
Other	any	4.2	0.4	0.1
Total	*All opportunistic*	**1132.1**	**100.0**	**35.3**

From 2019 to 2020 HPCs **contribution had x3 increases**

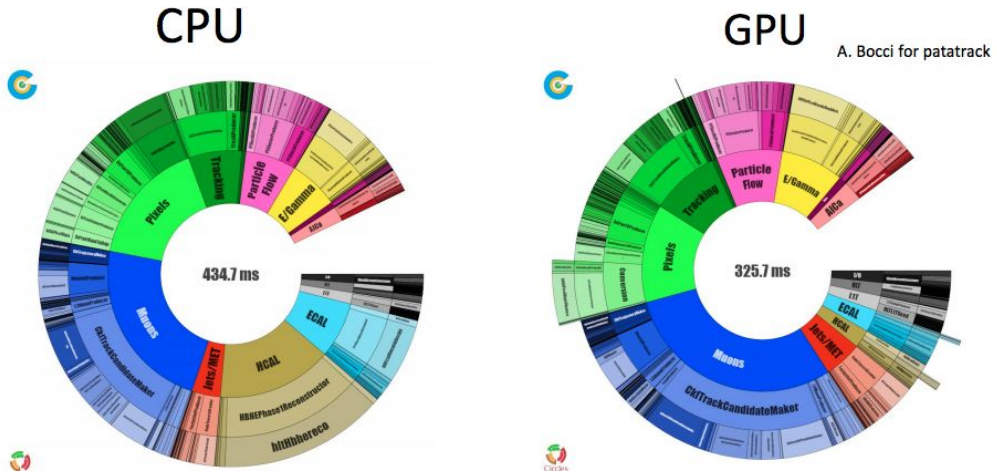
- From 33.6 to 108 kHS06/y.

GPUs and heterogeneous computing

- Since 2019 (Patatrack) CMS has started woking the porting the code to CUDA
  - Currently offloading about 25%
- Currently investigating the adoption of High Level Frameworks such as Alpaka, Kokkos and SYCL
  - allows writing single common code basis
  - EuroHPC/TEXTAROSSA Project
    - CNAF and PISA involved

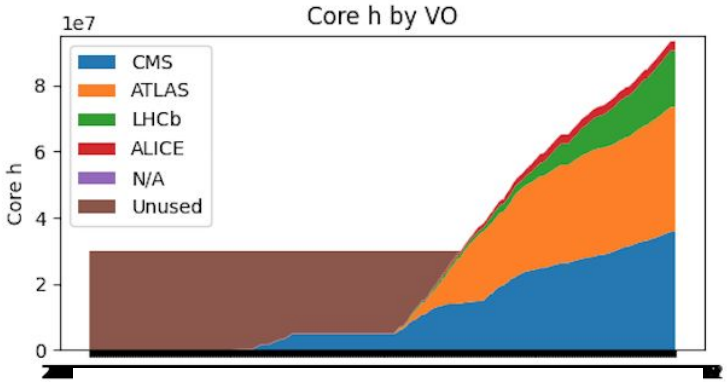
25% CPU time reduction when offloading pixels, ECAL and HCAL to GPU

(GPU used at 25%)



See S.Dal Pra @ ISGC

A few words on our national (INFN) landscape



PRACE Project Access. PI T.Boccali. 30McoreH

Integrating Tier1-CNAF and CINECA in the most transparent way to the end user/ experiment

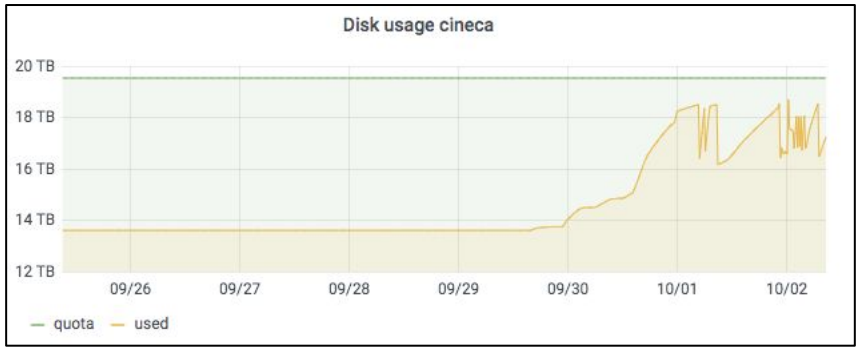
External networking enabled to **CNAF and CERN**

- Partial routing to CNAF storage / squids over the dark fiber @ **40 Gbit/s** (will improve with next machine)

Enough to guarantee access to CNAF storage and conditions for experiments' workflows + HTCondor pilot communication

Moreover: We use it also for accessing external data via proxies (xrootd)

A **HTCondorCE/Slurm** allowed on one CINECA login nodes



Looking at ahead: Marconi 100

Nov20 top500.org

Rank	System
11	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
Quick startup guide

Started working on the integration of the Marconi100 at CINECA (Power9+GPU), INFN got 3.5 MCoreH for 2021

- Enable **multi-arch support** for CMS production/analyses jobs
- Perform the **physics validation** on Power9 for its utilization in physics analyses

Currently:

- CMS software stack fully ported to POWER9
- Established a complete integration stack of the CMS Workload Management (WMAgent and CRAB)
- Successfully (technically) tested analyses and Release Validation workflows

GPUs:

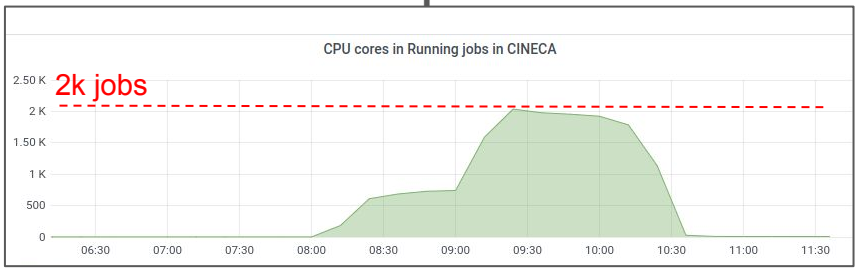
"Performance" in A.U. of the HLT application (the higher the better)

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

A first "scale" test performed as well



This first attempt is really promising, the focus now is on

- setup consolidation
- larger production

A few more words on HPCs with no-network

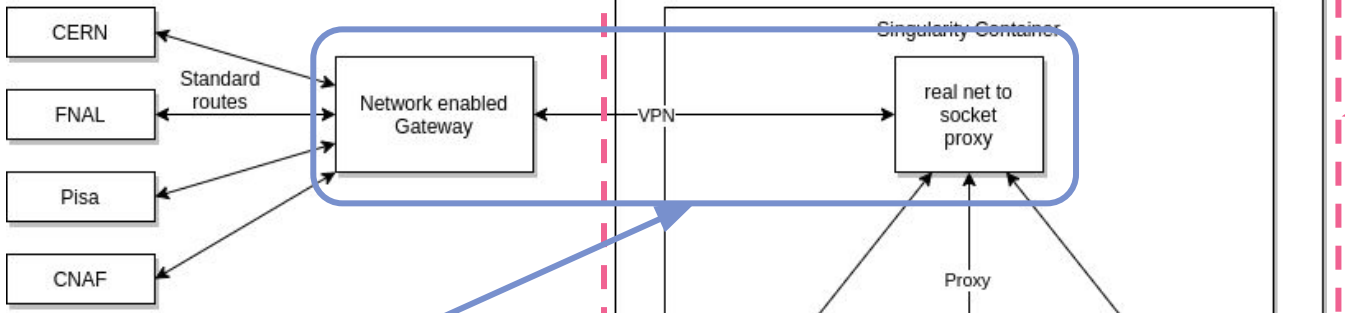
- **Negotiate** with the site for at least **minimal connectivity** (as for CINECA: ok for CERN and CNAF)
- **Encapsulate network** on something else (for example BSC(*): via FS, but it needs a service by service approach...)
- Deploy **large containers** to avoid needing CVMFS/Conditions (but condor management traffic? Stageout? Input data?)
- **Create edge services** to transmit / encapsulate / translate accesses
  - A Squid is (also) such a thing
  - ATLAS Harvester (**) is (also) such a thing
  - ArcCE , DIRAC
  - An xrootd proxy is (also) such a thing
  - How many services do we need to bridge one by one? (xrootd, srm, webdav, condor connections, dashboard reporting, DBS, DAS, Frontier, voms connections …)
- We need a lower level solution, **below the application layer**

- Having routing by kernel to a NAT is clearly a solution (but it needs the site admins to agree/implement)
- Having a VPN from the site to “somewhere else” is a solution (but it needs the site admins to agree/implement)
- If the sysadmins agree with deploying the previous two, the problem is not existing…
- … but difficult to find the agreement in some (known and future) cases

*: Exploiting network restricted compute resources with HTCondor: a CMS experiment experience
**: Harvester : an edge service harvesting heterogeneous resources for ATLAS

See M.Mariotti @ISGC

The overall idea



Isolated node

"You shall not pass"

… and something that converts an existing routed connection into a proxy socket.

Tsocks

After some research, we came up with the idea of wrapping (PRELOADING) the net calls to a local socket...

Two possible solutions:

1. **ssh -D**
2. **tunsocks + openconnect**

Summary

- HPC will play a key role in the future computing of the Experiments
  - Requires both technical and political effort

- Although many progresses have been done by all the experiments, not all the workloads can't be efficiently run on HPCs
  - Suitable for opportunistic computing but **we are still not ready for replacing grid-based resources** (pledges) with HPCs

- A lot of work still needed to deploy experiment software to fully profit from the accelerators and variety of processors at HPCs

- At national level, INFN the investment done at MarconiA2 with (Grant LHC) is starting paying back with ongoing Marconi100 system
  - in turn will be of fundamental importance for Leonardo, the pre-exascale system
  - Moreover it is a valuable experience in view of the new technopole datacenter under construction

Crediti per il lavoro di integrazione CNAF-CINECA, contributi ad R&D, supporto etc.

- Stefano Dal Pra
- Stefano Zani
- Gaetano Maron
- Luca dell’Agnello
- Lucia Morganti
- Daniele Cesini
- Vladimir Sapunenko
- Mirko Mariotti