Analisi e la gestione dei Big Data in ambito multidisciplinare: Data Lake e Cloud

Daniele Spiga - INFN Perugia
On behalf of

INFN
- Diego Ciangottini, Sara Cutini, Matteo Duranti, Mirco Tracolli, Pasquale Lubrano

UniPg
- Mirko Mariotti

PhD
- Giulio Bianchini, Tommaso Tedeschi

Outline

Introduction: The High Energy Physics Computing Context
- A quick look at national landscape

The technical challenges ahead of us

An high level overview of the **local activities**:

- **HEP and beyond**

Summary

spiga@pg.infn.it

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**
- From these needs many technological changes derived, main examples:
  - **The GRID**
    - 161 official sites, in 42 countries
      - "We only miss Antarctica"
    - Pledged resources
      - ~ 8 MHS06 (~800kCores)
      - ~700 PB disk
      - ~1 EB tape
  - **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

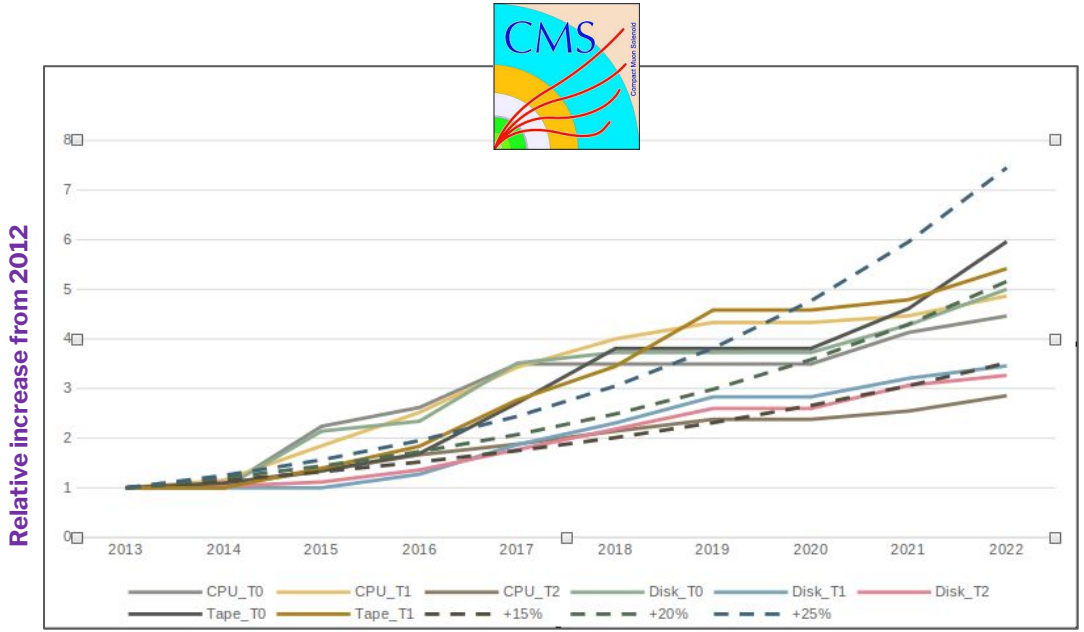
Nota: 1 PC attuale ~ 1 SI2k

IFAE Catania 2005

spiga@pg.infn.it

Let's focus on LHC (the biggest collider in operations)

- **LHC started its operations in 2009**
- By now, 2022, the needs have increased 5x 2007 → 2012 …
- … and then another 5x 2012 → 2022

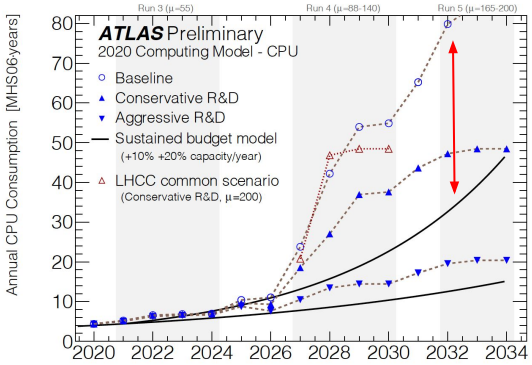


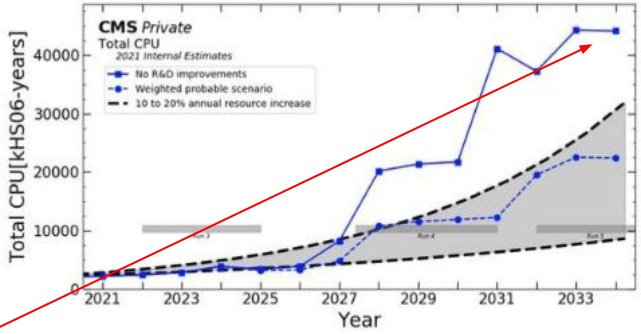
Computing (its cost!) IS the limiting factor of experiments @ LHC: if we were able to spend 10x in that, or do things 10x more efficiently, we could do more physics

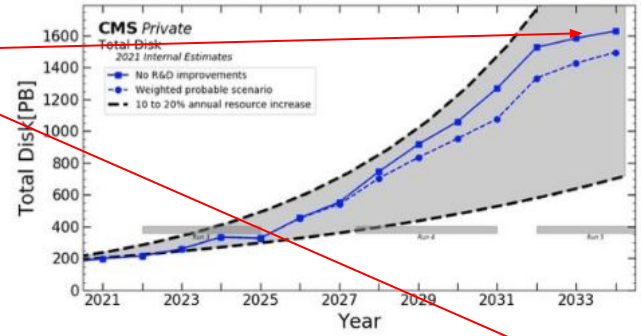
spiga@pg.infn.it

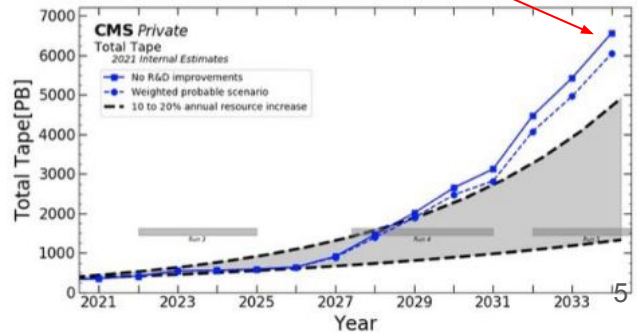
And what comes after LHC?

- Even without going to look into the computing of FCC, ILC, CLIC, CEPC (*), the next 10 years are going to be problematic.
- High Luminosity LHC: 6x more complex events, 10x more events collected by the experiments, 10+ years of operations
- Consequence on computing is “huge”: ~10x increase in computing needs and storage .. And no increase in money!









So… what to do?

The mantra is: to propose new ideas, many R&D activities and to implement testbeds …

- **Heterogeneous Architectures** (GPUs etc)
  - Direct impact on software and computing
- **Facilities integration**
  - Transparent exploitation of HPC, Cloud
- **New approaches to data analysis**
  - Facilities, data format, interactive vs batch
- Software optimization
  - Performances
- **Distributed Computing models**
  - DataLake, Cloud..
- **Machine Learning**
  - Reco, Analysis..

Today the focus is on all those activities where local effort is directly involved

spiga@pg.infn.it

… And what about INFN in this puzzle

INFN is a pioneer in the design and implementation of large-scale computing infrastructures and applications.

- Primarily developed to address the needs of the latest generations of high energy physics (HEP) experiments but **now, rapidly, extending to other communities.**

The current production system is composed by

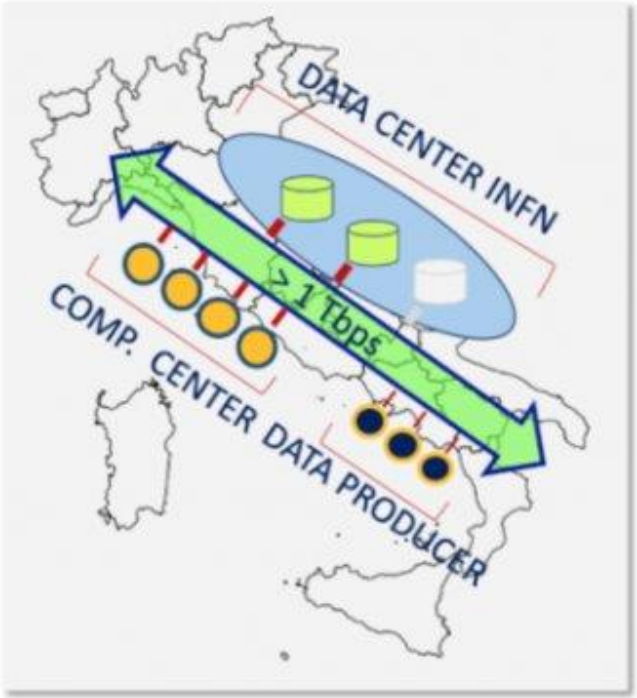
- **9 medium sized centers (known as Tier2 in the LHC Computing Grid)**
- **1 big center Tier-1, CNAF,** (located in Bologna)
  - ISO/IEC 27001, 27017 e 27018 certified to allow the management of data and applications in cloud, including sensible data such as medical and clinical one
- **A National federated Cloud: INFN-Cloud** (see later)
- All the INFN centers have been connected at 10-100 Bbit/s through the GARR network

Overall 100 **petabyte (PB) of storage and more than 100.000 CPU Cores.**

The INFN-Cloud Project

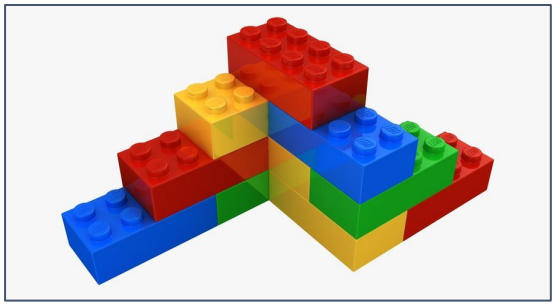
The INFN-Cloud project, launched at the beginning of 2020 and currently in production, is the driving force for the Cloud development of all INFN initiatives.

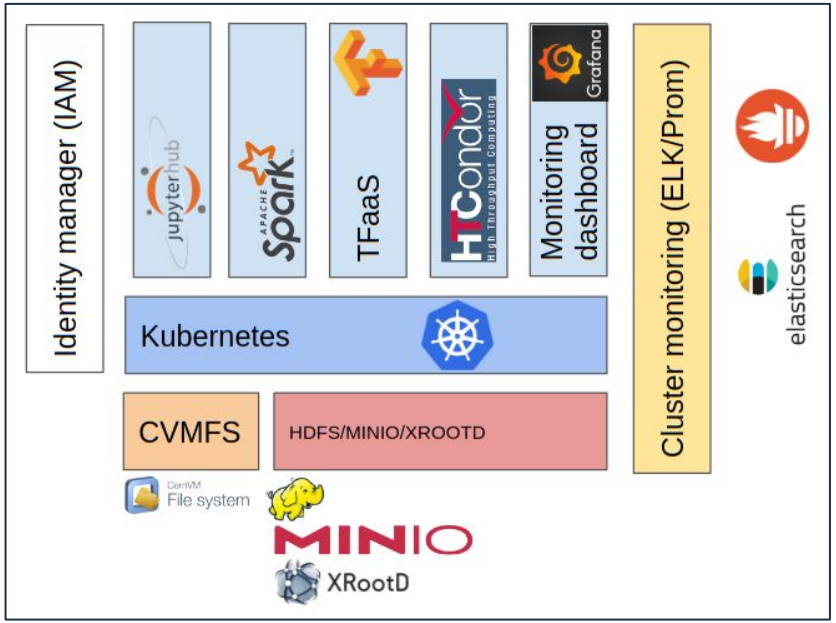
- **A multi-site federated Cloud infrastructure** owned by INFN, **expandable to other Cloud infrastructures and resources**
- **A set of services** that can be used through a portal, from a terminal or with a set of APIs.
- **A "high-level" mechanism for adapting and evolving the service portfolio according to the needs and requests of users.**
- A fully distributed intra-INFN organization for the support and management of infrastructure and services.
- A series of rules for access and management policies of INFN Cloud resources that incorporate INFN regulations and the more general national ones.



Cloud native solutions for scientific application

Lego blocks solutions fully integrated with INFN-Cloud portfolio of services





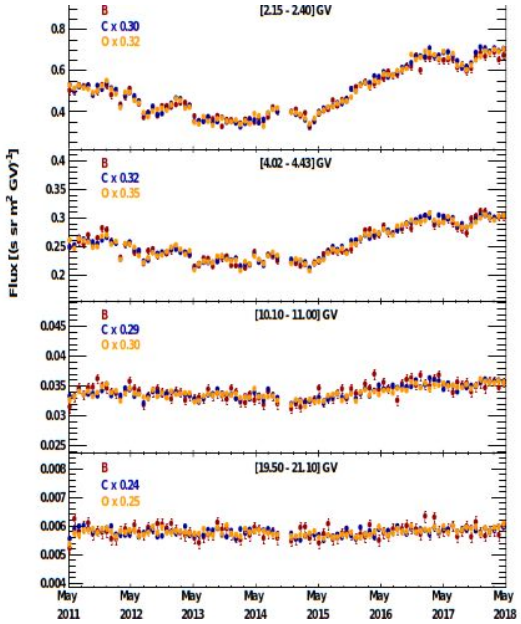
- **Highly Customizable** to accommodate needs from diverse communities –> **See later**
  - Built on top of modern industry standards

Porting AMS analysis in Cloud

This has been a demonstrator in the EOSC-hub (H2020) project

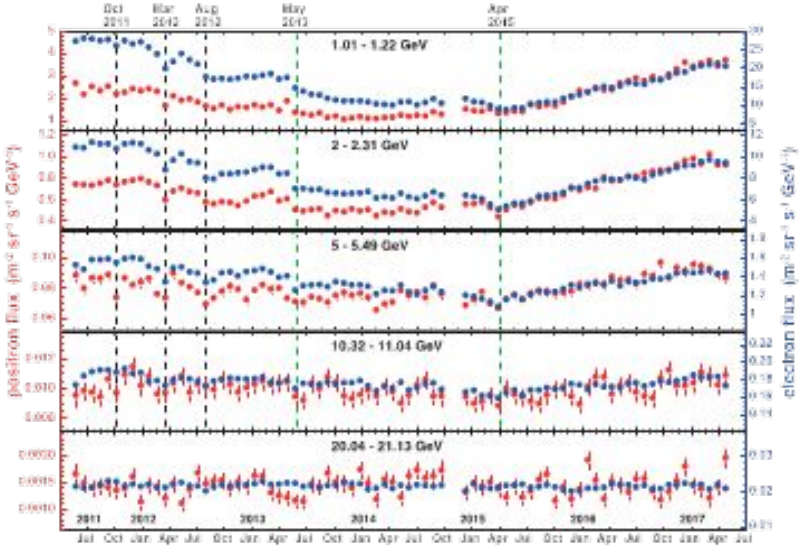
INFN Istituto Nazionale di Fisica Nucleare

- The Alpha Magnetic Spectrometer measures Charged Cosmic Rays (0.1 – 2000 GV) in space since 2011, May 19th
- The ~ 160 billion events (~ 180 k science runs), once reconstructed in ROOT format, weight ~ 1 PB
- The analyses are performed on stripped samples (i.e. streams) with a lighter data format (i.e. ntuples)
  - a job to produce a single run needs ~ 2 hr and produces O(102 MB) ntuples
  - every analysis target (e.g. electrons/positrons vs. ions) requires its own ntuples set



left - Boron, Carbon (scale = 0.24) and Oxygen (scale = 0.25) fluxes as function of time (Δt = 27 days)

bottom- Electron and positron fluxes as function of time (Δt = 27 days)



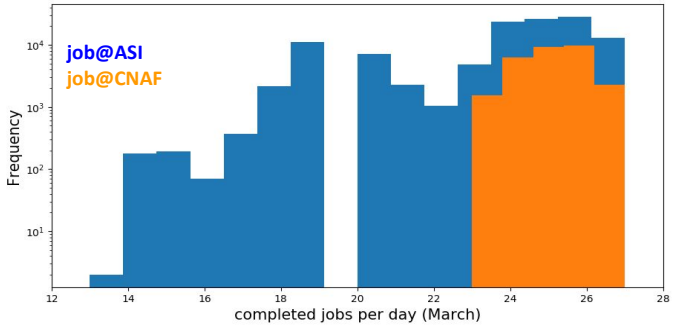
1. AMS collaboration previously published B, C and O fluxes only as a function of energy and time-integrated. **This new analysis, by INFN-RM2, has been performed using the ntuples produced running on DODAS;**
2. Electrons and positrons fluxes, as a function of time have been already published with 27 days time granularity. **A new analysis, by INFN-PG, and using the ntuples produced on DODAS, is extending the time range and producing the electron (positron) fluxes on a daily (weekly) basis;**

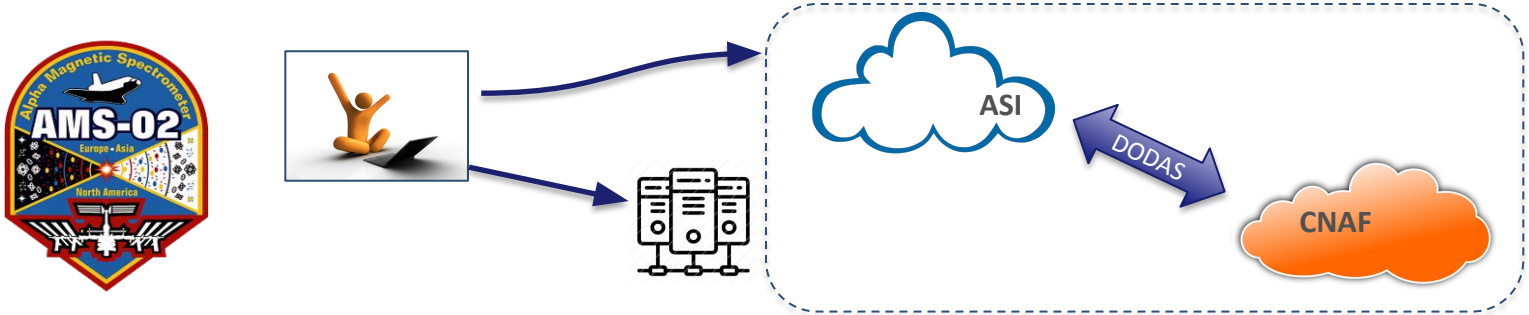
Exploiting resources federation: The ASI example

The Space Scientific Data Center (SSDC) of the Italian Space Agency (ASI) host an AMS farm.

- no experiment dedicated manpower
- no specific expertise on AMS software and computing environment

An example of Stateless Site Providers





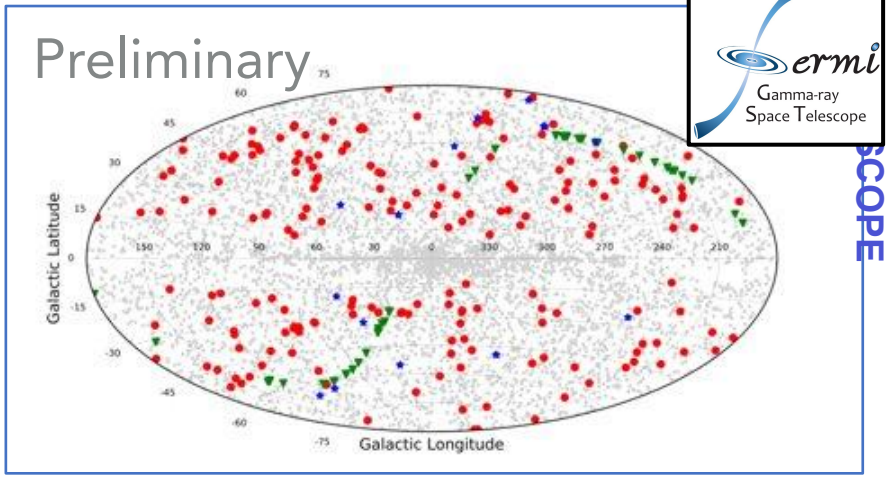
Started with a very specific target… now a generic solution, exploited by FERMI and will be functional to HERD etc etc…

Running FERMI-LAT analysis on Cloud resources

The LAT instrument onboard of Fermi gamma-ray science telescope (Atwood et al. 2009) observes the sky in the gamma rays range between 30MeV - 300 GeV since August 2008.

- Extract catalog of transient sources (monthly basis) from Fermi-LAT data (1FLT; Fermi-LAT collaboration in preparation)
  - 10 years of data in monthly timescale ⟶ 120 independent skies + 120 (15-day shifted month)
- Detected ~1000 seeds/monthly skies
  - ~260 binned maximum likelihood analysis (ML) for each month.

- **Submitted roughly 60k ML analysis jobs —>Very time-consuming! ~ 960 h of computing time without interruptions**



Aitoff projection of 1FLT in red, sun detections from transient catalog in green, and GRB detection in blue. Standard catalog (4FGLDR2) in gray

507 new detections —>extraction of standard products: monthly light curves **(120 ML jobs per source)** and Spectral energy distributions **(4 ML jobs per source)**

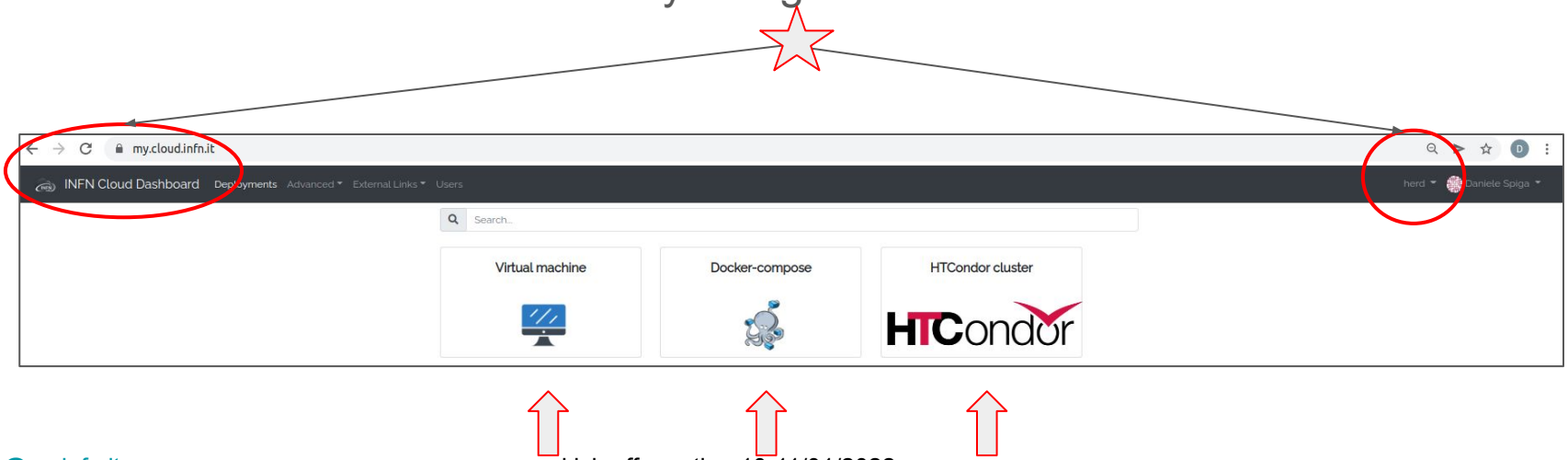
FERMI-LAT has been selected as use case in the context of **EGI-ACE EU Project** (S.Cutini PI)

FERMI GAMMA-RAY SCIENCE SCOPE

What is in the pipeline: The HERD Experiment

A nice example of how most of what have been developed **can "smoothly" evolve in order to support diverse experiments and their computing needs**

HERD integration within INFN-Cloud infrastructure already started and currently being commissioned



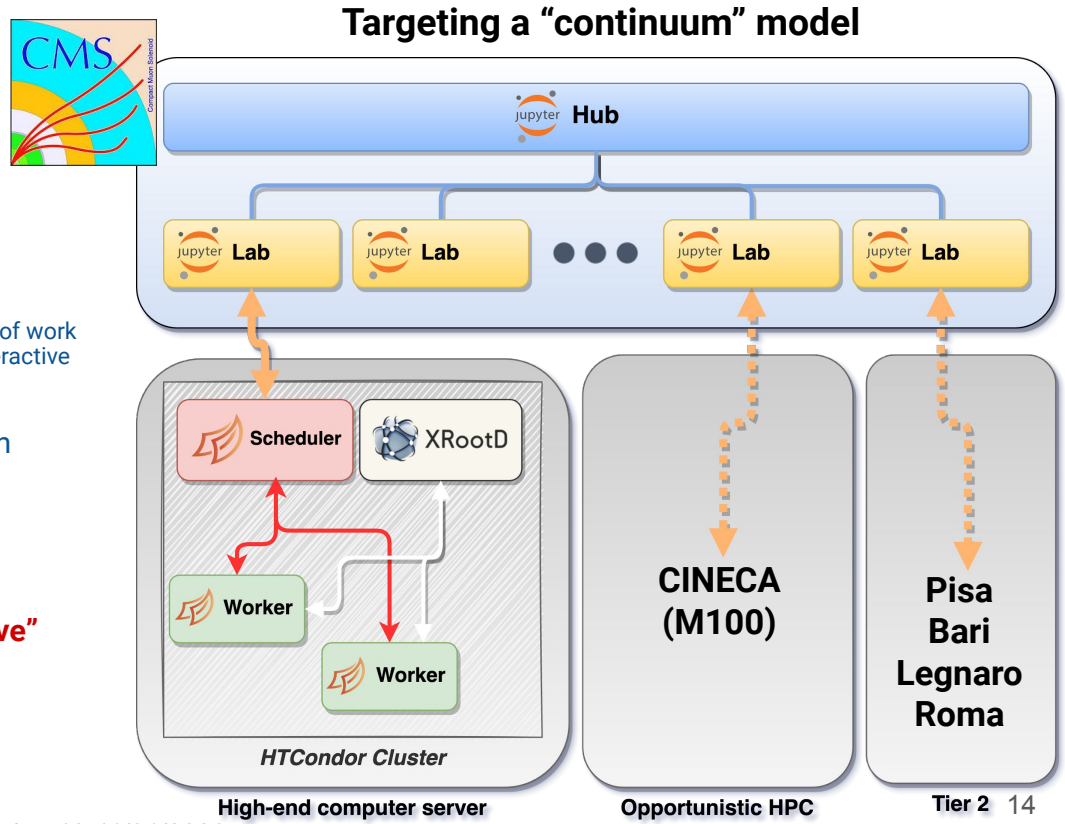
“Big Data”: solutions for Interactive Analysis

Building on “de facto” standards

- **Interfaces**
  - **JupyterHub** as user entrypoint
  - **JupyterLab** to manage the user-facing interface
  - Direct access to HTCondor
  - User interface (either from JLab or old fashoned UI)
- **DASK** to introduce the scaling over a batch system
  - Multiple clusters per user → DASK cluster as atomic unity of work
  - With some caveat they can be seen like the CRAB task interactive equivalent
- **HTCondor** as the batch system of choice
  - User prioritization and in general configuration tuning is under study
- **XRootD** as data access protocol toward AAA:
  - Here we foresee the usage of caching layers (see later)

For the whole chain of software we kept it as **a “token native” system based on IAM@CMS**

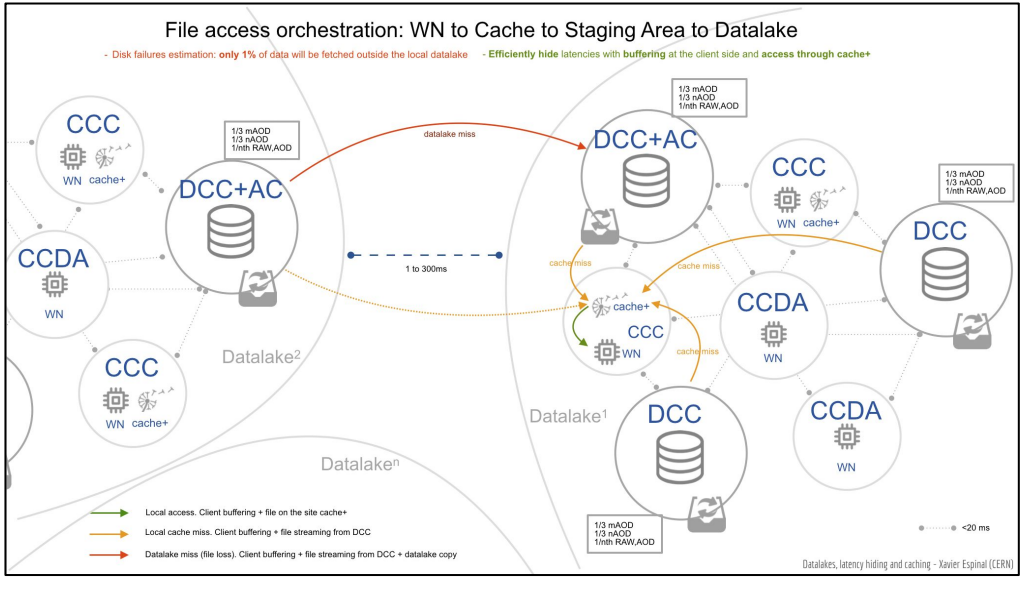
And most importantly: **everything is just a “Lego brick”**



The Data Lake in the context of WLCG (HL-LHC)

- Optimization is needed to **reduce the hardware usage and operational costs**
  - No a one size fits all solution

From Data Lake Straw Proposal @ DOMA



Terminology:

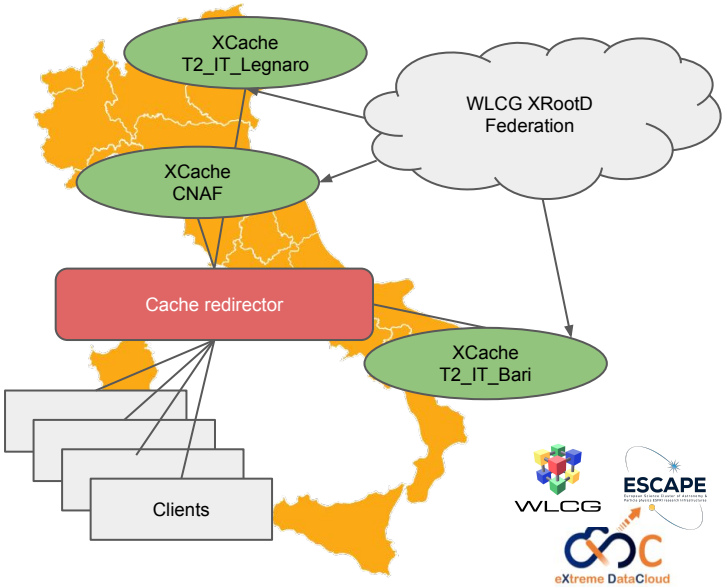
AC - Archive center Defined as Tape or tape-equivalent-QoS enabled center able to archive custodial data.
DCC - Data and computer center providing disk-equivalent QoS storage
CCC - Compute center with cache
CCDA - Compute center without cache: relies on accessing all data via the network from either a CCC or a DCC

A small number of Data Lakes across the world

- Reduced number of storage endpoints wrt the current WLCG model

Envision a mix of **distributed caches** directly accessed from compute nodes

The INFN sites as nodes of a Lake: a CMS Testbed

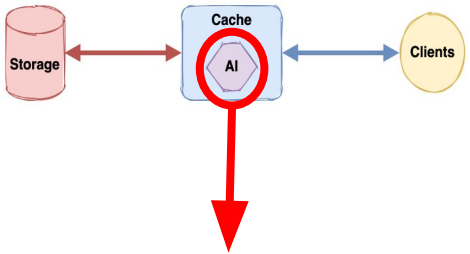


- **CNAF XCache redirector federating 3 XCache servers**
  - CNAF server (5TB spinning)
  - Bari server (10TB gpfs)
  - Legnaro server (22TB spinning)

Smart cache system using AI algorithms

Classic caching approaches (LRU, LFU, etc..) are not able to address the efficient caching problem
They do not adapt to changes and do not have any predictive feature
A smart approach is needed (with predictive ability: thus Machine Learning-based)

AI directly manipulates cache memory, deciding what to write or delete



Collaboration with Dip.Matematica e Informatica
- **Valentina Poggioni**
- **Marco Baioletti**

QCACHE:
a Reinforcement Learning-based framework

Integrating heterogeneous resources: Marconi100 at CINECA

What:

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

Why it is of interest:

- Provides access to a different platform, which could be used in next generation HPC systems
- Provides access to GPU deployment (Nvidia V100)

Would allow to

- Demonstrate we can integrate and use non x86 platforms
- to perform physics validation on Power9, and bless the platform for production in CMS
  - **LHC-Italy** got 3.5 MCoreH (~ 20 nodes) [**2021**]

Aside notes:

- Since 2021 CINECA grants are only for PowerPC
- M100 and Summit have basically the same architecture

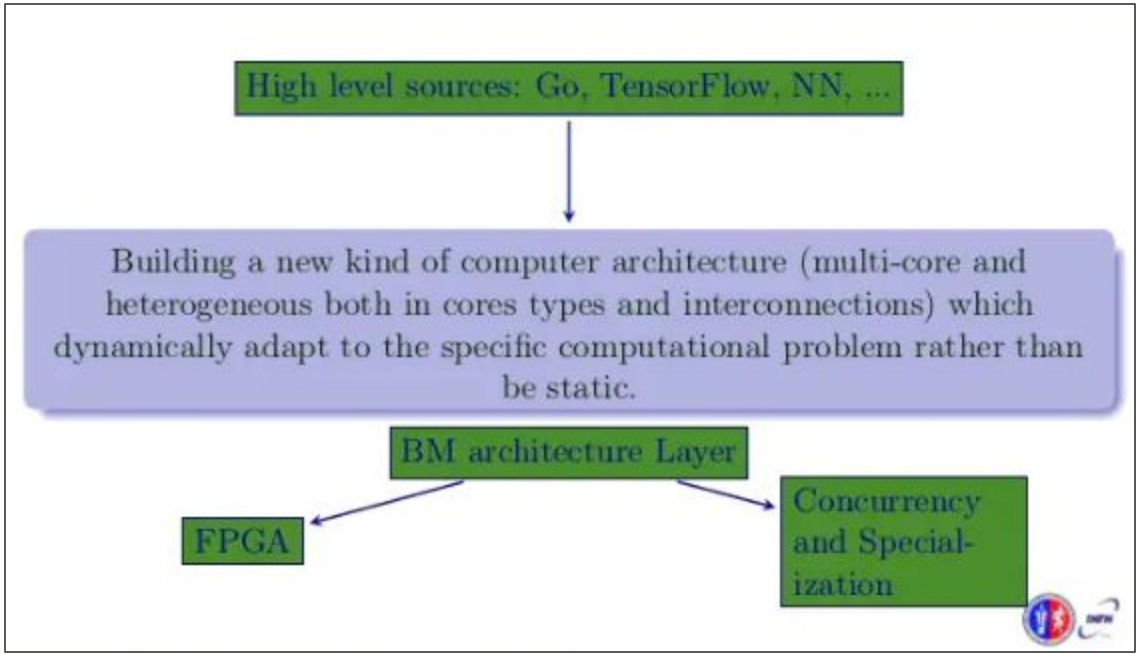
Prototyping solutions for heterogeneous platforms

The aim is to exploit FPGA in scientific computing

Developing software layers to enable easy access to

- Massive parallelism
- Porting legacy application/software

PhD - PON



The Bond Machine project [PI Mirko Mariotti]

Beyond physics data analysis: Heterogeneous Data

Expertise developed to support HEP experiment data analysis are a key to contribute in other domains:

Data Analysis

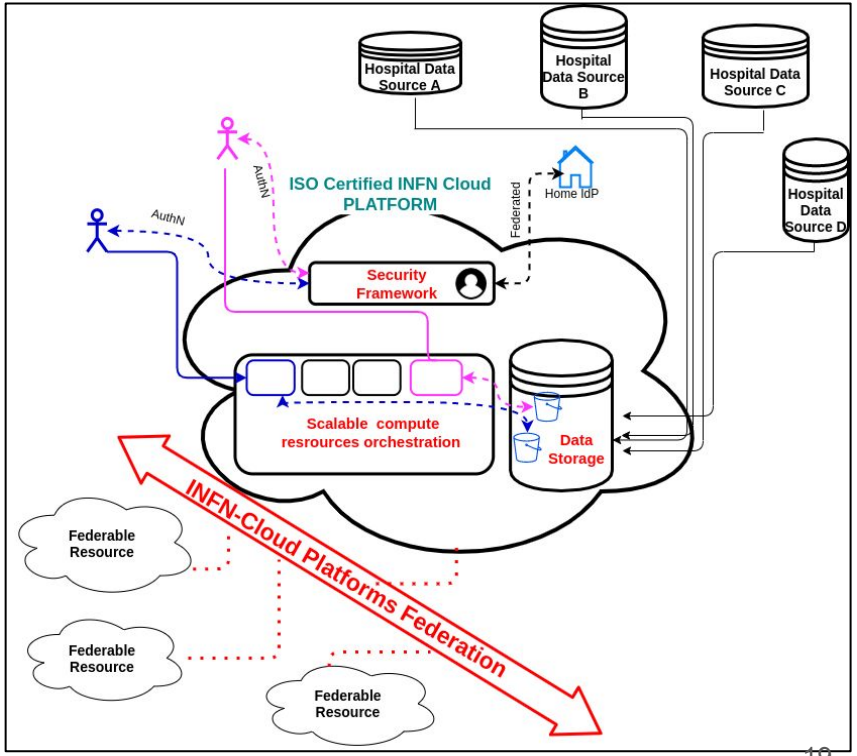
- To develop models and to implement statistical data analysis

Data Curation and Data Management

- Organization and integration of data collected from heterogeneous sources; Enabling FAIR (Findable, Accessible, Interoperable, Reusable) data repositories

Integrated and Certified Computing Infrastructure

- A ISO 27001 / 27017 Certified Data-Lake to manage confidential data (ISS, Hospital, ASL)
- An easy to use computing platform fully integrated with the **INFN-Cloud national infrastructure**



spiga@pg.infn.it

The PLANET Project

Collaboration with Dip.Medicina
- **Prof. Giuseppe Ambrosio**
- **Prof. Fabrizio Stracci**
- **Prof. Giampaolo Reboldi**

INFN
Istituto Nazionale di Fisica Nucleare

An observational (ecological) study to evaluate the association between air pollution and Covid19, taking care of a variety of components that are supposed to influence rates of SARS-COV-2 diffusion and infection

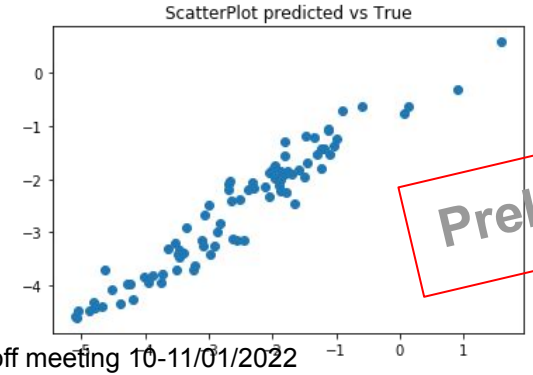
- A synergy between INFN and epidemiological and medical knowledge (Univ. of Perugia)
- Nicely progressing toward the established objectives

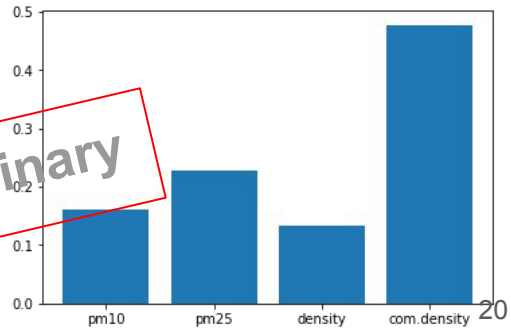
Current focus on Feature importance evaluation: assign a score to input features based on how useful they are at predicting a target variable (Covid19)

- **Models such as: Random Forest; k-nearest neighbors;**

Heterogeneous data sets

- **Covid19:** ISS, Hospitals, ASL
- **Pollutants**: Copernicus (CAMS) and ARPA
- **Meteo**: Copernicus
- **Population:** ISTAT
- **Socio Economics**: Deprivation Index - Commuting density





Preliminary

Summary

Historically the R&D operations on computing has been motivated by the needs the big HEP Experiments

- Nowadays the paradigm shift demonstrates how developed solutions are generic enough to support a wider range of requirements…
- **Expertise and competences really matter**! A key to establish successful synergies

Locally: very active contribution to the national and international landscape

- **CMS (HL-LHC), AMS, Fermi** (**HERD** is ongoing) ... More will come **(i.e. The Einstein Telescope [ET])**
- Achieving expertise also in data treatment beyond the physics: **mainly clinical data**
- Fruitful collaboration with other department @ UniPG already established
  - Consolidation process is foreseen
  - In this context: currently building with UniPG a **EDIH proposal** for a hub@Umbria : **U**mbria **D**igital **D**ata (**UDD**)

spiga@pg.infn.it

Backup

spiga@pg.infn.it

Cache effect on CPU efficiency @ CMS

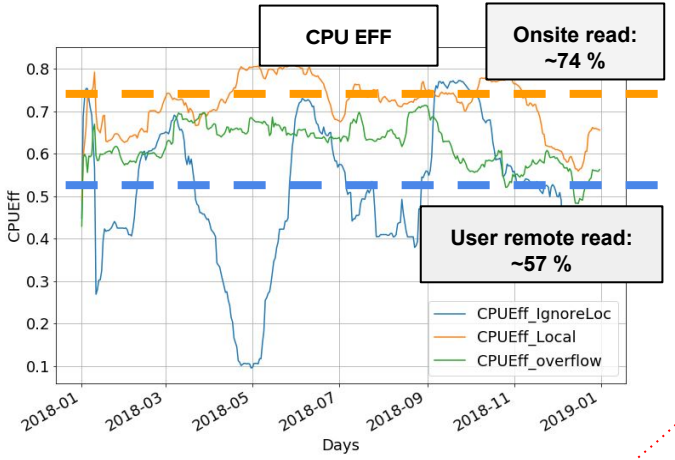
- We studied monitoring data of the whole **2018 CMS analysis workflows**
  - Remote data read costs on average **about 15% of CPU time w.r.t. Onsite data reading**

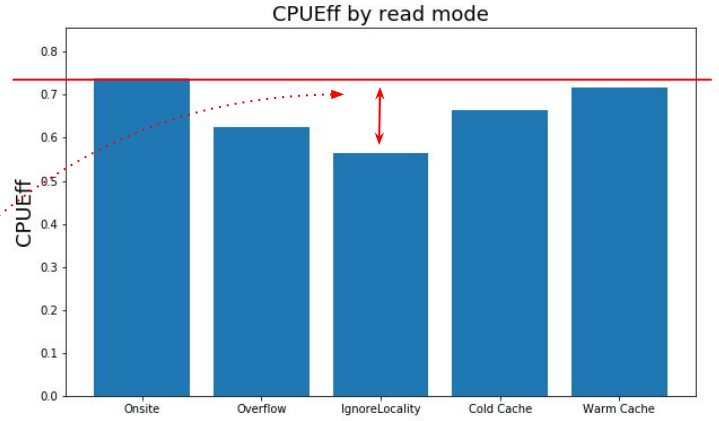
CPUEff = sum(cpu time) / sum (job time)

From data@CERN MONIT hdfs

Measured on testbed

@Italian Tier2's

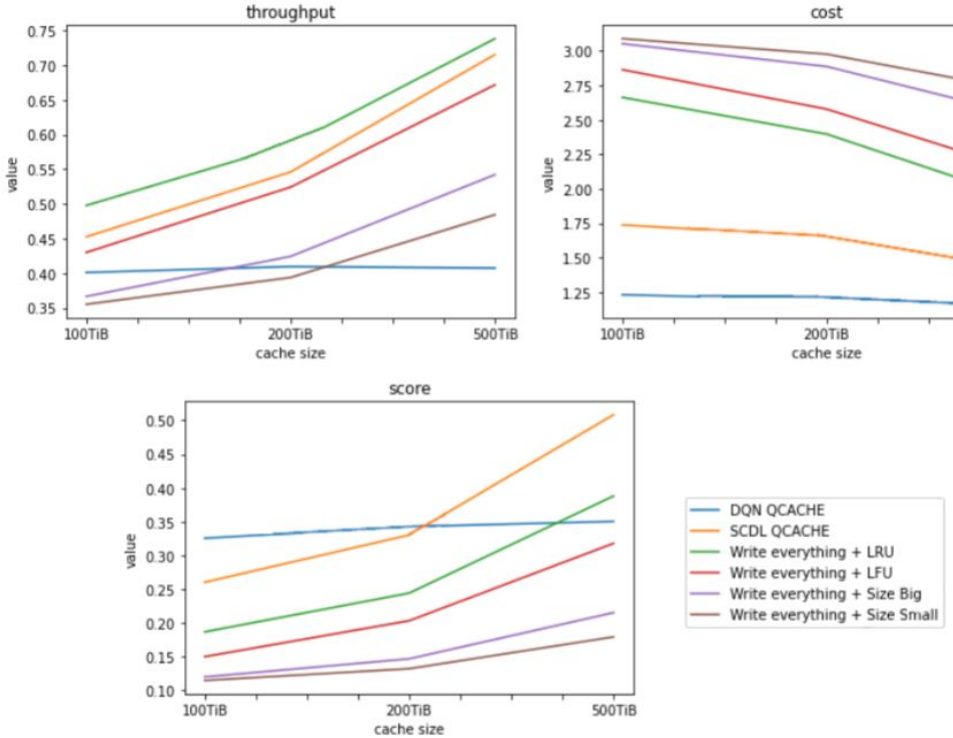




Caches allow to reduces the overall WAN traffic and, makes the processing job that requested the data **more efficient** by reducing I/O wait time for remote data.

Smart Caches: Results

Daily values averaged across the year



100 TiB

Algorithm	Score	Throughput	Cost
DQN QCACHE	**0.33**	0.40	**1.23**
SCDL QCACHE	0.26	0.45	1.74
Write everything + LRU	0.19	**0.50**	2.66
Write everything + LFU	0.15	0.43	2.86
Write everything + Size Big	0.12	0.37	3.05
Write everything + Size Small	0.11	0.36	3.09

200 TiB

Algorithm	Score	Throughput	Cost
DQN QCACHE	**0.34**	0.41	**1.20**
SCDL QCACHE	0.33	0.55	1.65
Write everything + LRU	0.24	**0.59**	2.40
Write everything + LFU	0.20	0.52	2.58
Write everything + Size Big	0.15	0.42	2.89
Write everything + Size Small	0.13	0.39	2.98

500 TiB

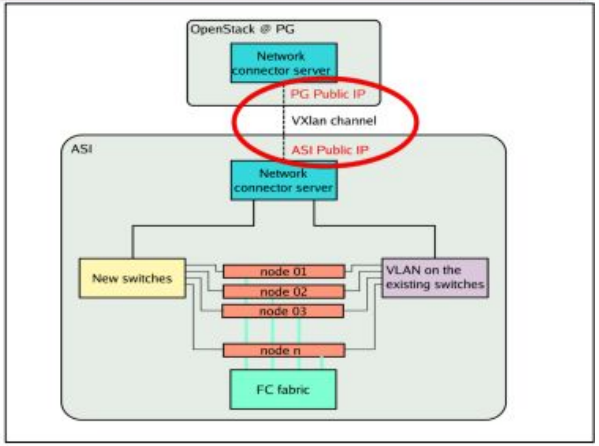
Algorithm	Score	Throughput	Cost
SCDL QCACHE	**0.51**	0.72	1.41
Write everything + LRU	0.39	**0.74**	1.90
DQN QCACHE	0.35	0.41	**1.16**
Write everything + LFU	0.32	0.67	2.11
Write everything + Size Big	0.22	0.54	2.52
Write everything + Size Small	0.18	0.48	2.70

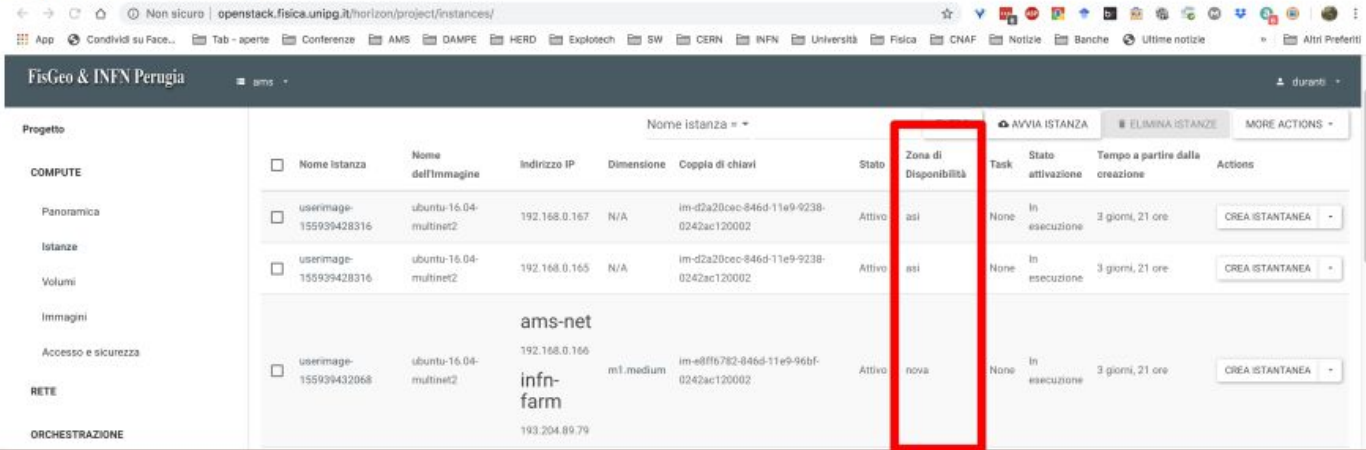
OpenStack@PG + ASI "zone"

At the Space Scientific Data Center (SSDC) of the Italian Space Agency (ASI) we have an AMS farm:

- 384 cores
- 90TB

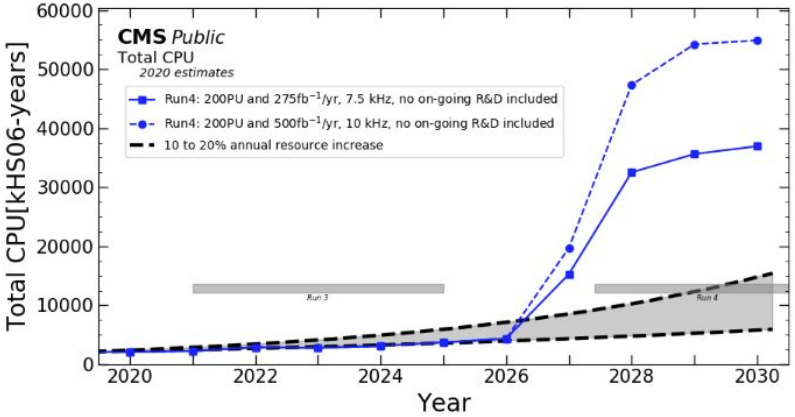
now (in part) managed by the OpenStack@PG

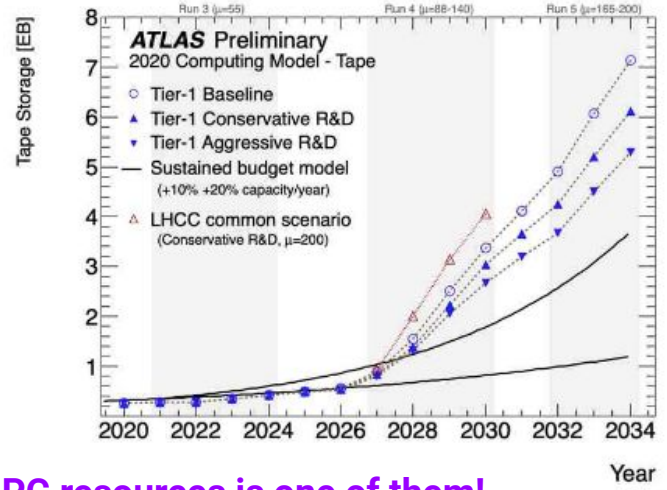




What is in front of us?

- 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

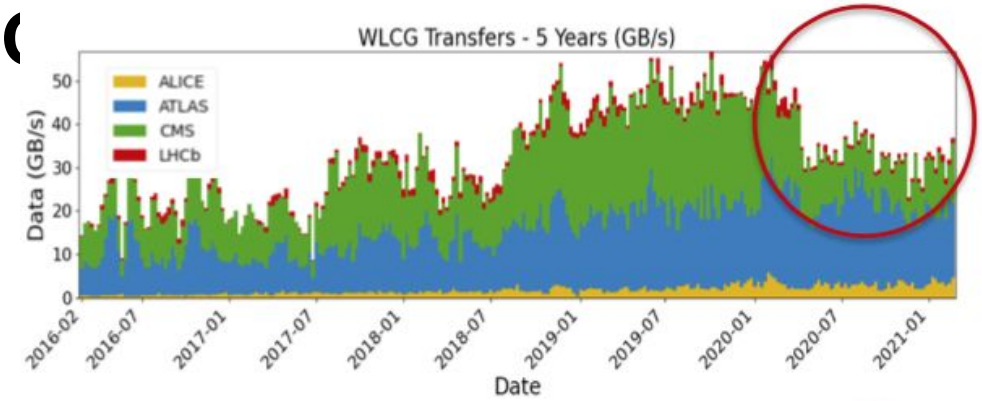




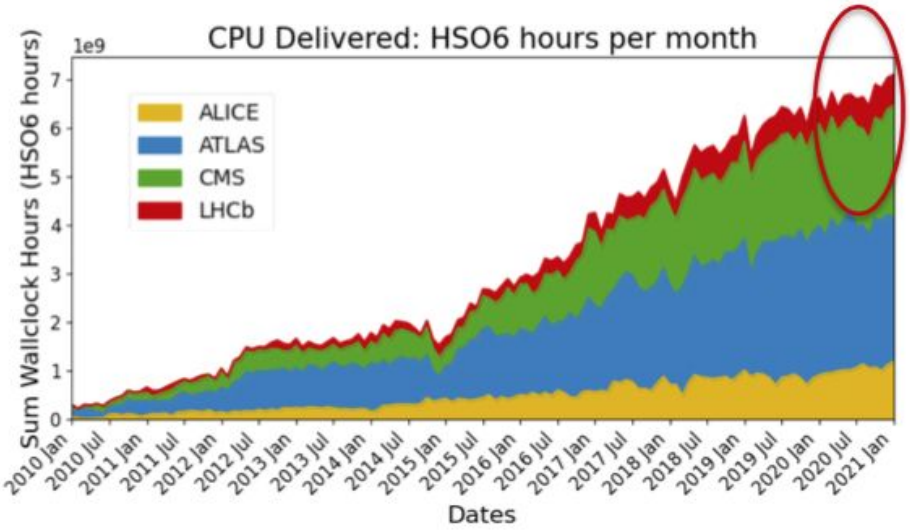
Many solutions / ideas under test; using HPC resources is one of them!

2021: WLCG (The (

- 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

I “servizi” per l’analisi dei dati

Il **portafoglio dei servizi Cloud dell'INFN** si basa su software open-source e su standard de-jure o de facto, seguendo il principio della composizione dei servizi

- ovvero procedure che consentono di realizzare la migliore soluzione dato un certo problema, utilizzando una **composizione di soluzioni di base**
  - **Semplificare e democratizzare** l’accesso a calcolo,
  - **Personalizzazare** le configurazioni,
  - **Estendere e comporre** i servizi in base alle nuove richieste