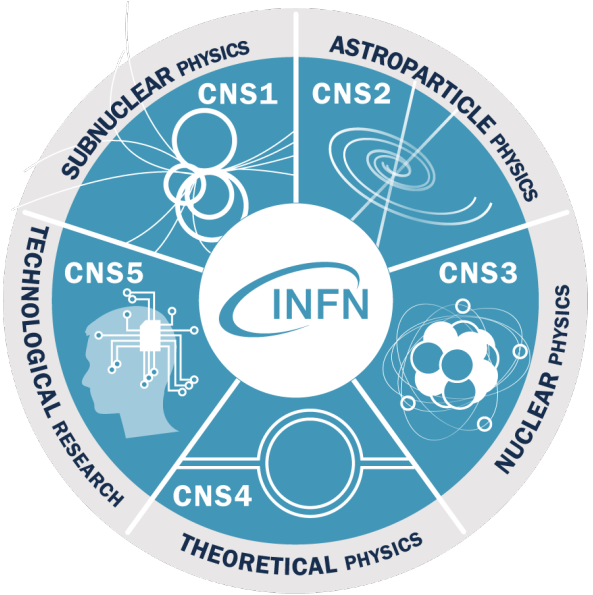
INFN and the evolution of distributed scientific computing in Italy

Federica Fanzago on behalf of the DataCloud INFN team

Credits to D.Salomoni for materials

The INFN mission

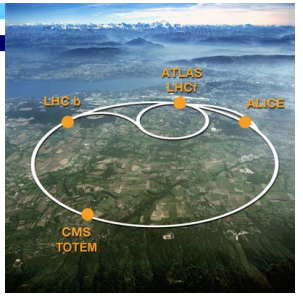
- “The National Institute for Nuclear Physics (INFN) is the Italian research agency dedicated to the study of the fundamental constituents of matter and the laws that govern them, under the supervision of the Ministry of Education, Universities and Research (MIUR)”.
  - Research activities are undertaken within a framework of international competition, in close collaboration with Italian universities.
- Research activities are organized in scientific areas
  - each one coordinated by a National Scientific Commission.



Facilities at INFN

- Distributed and collaborative work environment
  - within a network of collaboration, at foreign and international laboratories and at its own facilities in Italy.
- It was clear since the beginning the importance to own, operate and manage a reliable computing infrastructure.

LHC and Grid era



- INFN was one of main promoters of the GRID project to address LHC computing needs. Since then INFN has been participating to WLCG that includes more than 170 sites around the world, loosely organized in a tiered model.
  - In Italy, there are the Tier-1 at CNAF, Bologna and 9 “Tier-2” centers.
- INFN has been running for more than 20 years a distributed infrastructure which currently offers about 150000 CPU cores, 120 PB of enterprise-level disk space and 120 PB of tape storage, serving more than 40 international scientific collaborations.
- All the INFN centers are connected through 10-100 Gbit/s dedicated links via the GARR network.
- INFN manages and supports the largest public computing infrastructure for scientific research spread throughout the country.

From local experiences to INFN Cloud

- To support and evolve use cases that could not easily exploit the Grid paradigm, for many years several INFN sites have been investing in Cloud computing infrastructures
  - heterogeneous in hardware, software and cloud middleware.

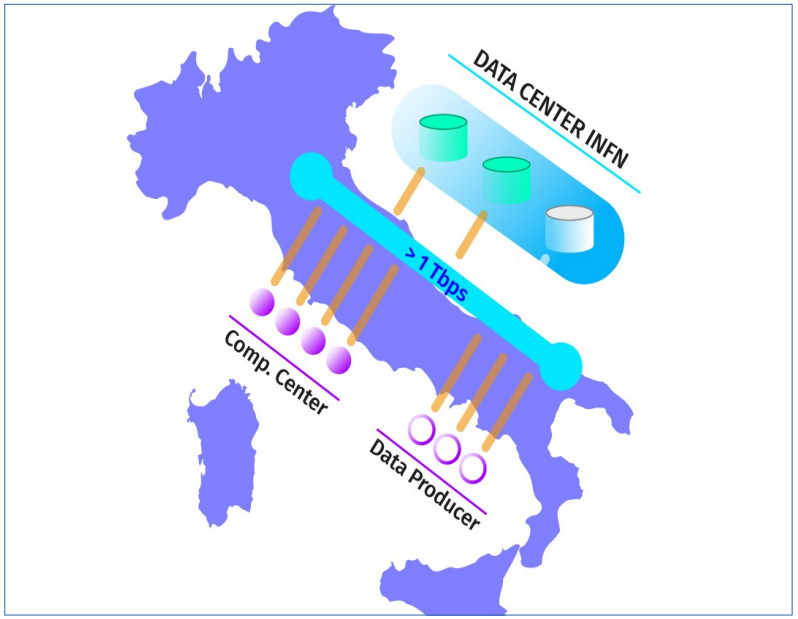
Cloud at CNAF
Cloud at Bari
CloudVeneto
Cloud at Torino...

- To optimize the use of available resources and expertise, INFN decided to implement a national Cloud infrastructure for research
  - as a federation of existing distributed infrastructures extending them if necessary in a transparent way to private and commercial providers;
  - as an "user-centric" infrastructure making available to the final users a dynamic set of services tailored on specific use cases;
  - leveraging the outcomes of several national and European cloud projects where INFN actively participated.
- INFN Cloud was officially made available to users in March 2021.

Federation of data, resources and knowledge

https://www.cloud.infn.it

INFN Cloud implementation

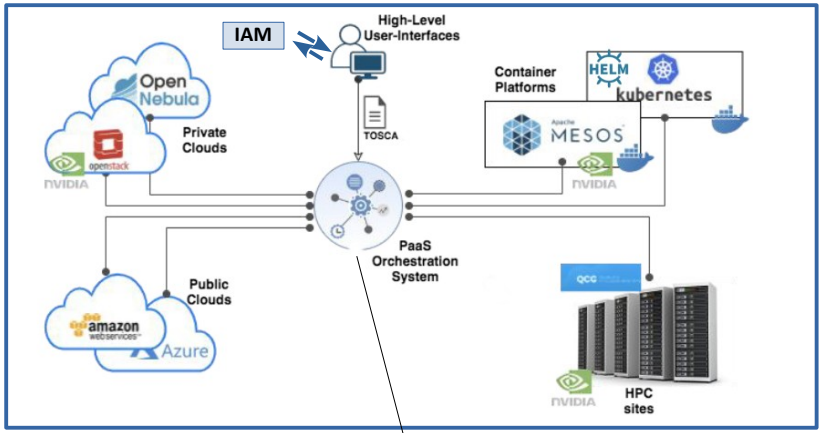


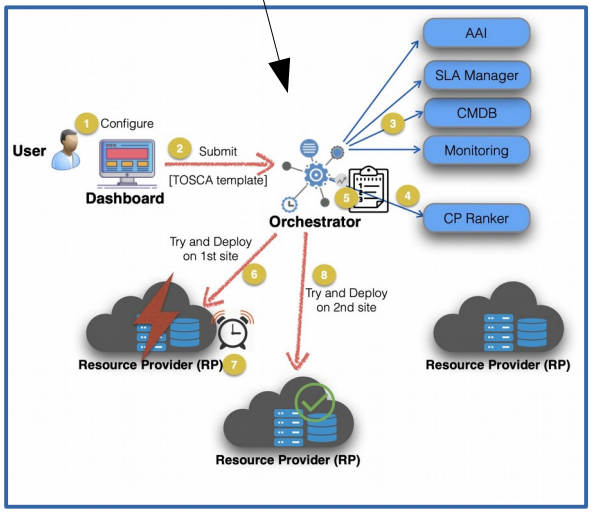
- The infrastructure is based on a core backbone connecting the large data centers of CNAF and Bari and on a set of loosely coupled distributed and federated sites connected to the backbone
  - backbone's sites are high speed connected and host the INFN Cloud core services.
- A site can join the INFN Cloud infrastructure accepting the Rules of Participation and after the approval of the INFN Cloud project management board.
  - Rules define access to resources and policies, according to INFN national and European laws.

- INFN Cloud's distributed organization provides support and management of both infrastructure and services.

Architectural key points

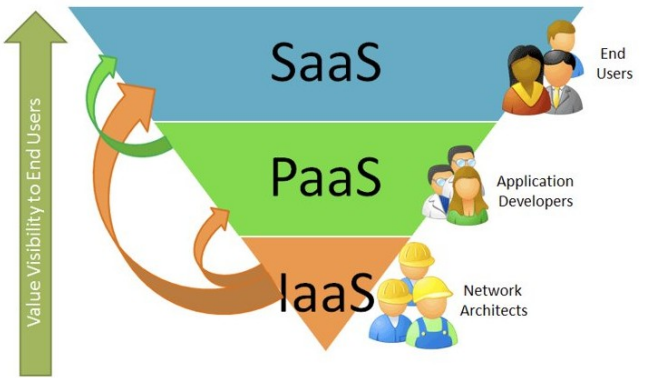
- Open source, vendor neutral architecture.
- Dynamic orchestration of federated resources
  - via the INDIGO PaaS Orchestrator across all participating Cloud infrastructures, according to agreed SLAs and Rules of Participation.
- Consistent authentication and authorization technologies and policies at all Cloud levels
  - via OAuth and OpenID-Connect, supporting also legacy AAI solutions, via INDIGO-IAM (Identity and Access Management).





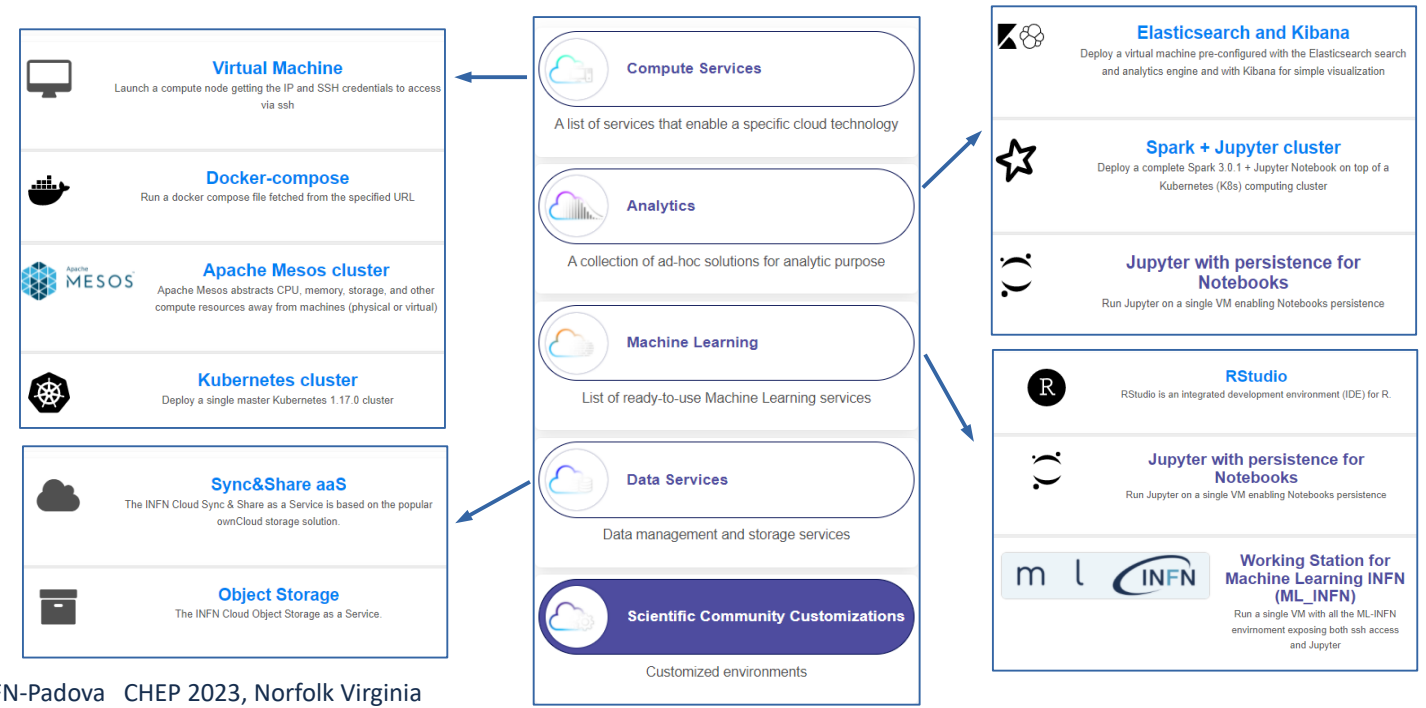
INFN Cloud portfolio

- INFN Cloud helps researchers in their daily analysis work that requires ever more complex workflows and computing knowledge.
- It provides a customizable and extensible portfolio of services
  - computing and storage services spanning the IaaS, PaaS and SaaS layers, with dedicated solutions to serve special purposes, such as ISO-certified regions for the handling of sensitive data.
- Services are instantiated through TOSCA templates and implemented using the “lego-like” approach, building on top of reusable components.



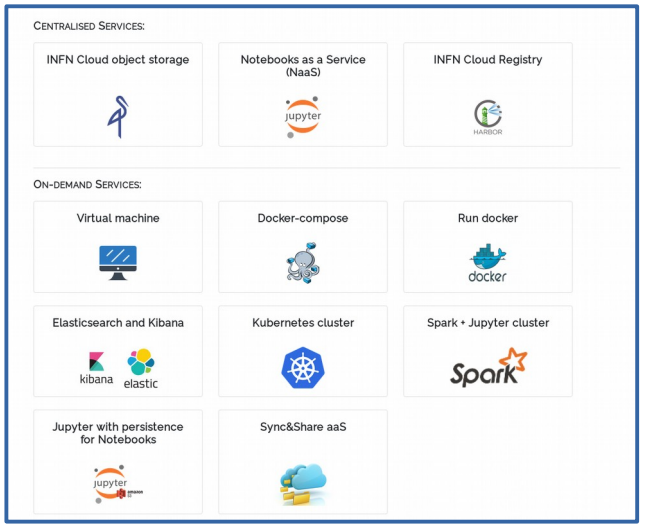
The dynamic catalog of services

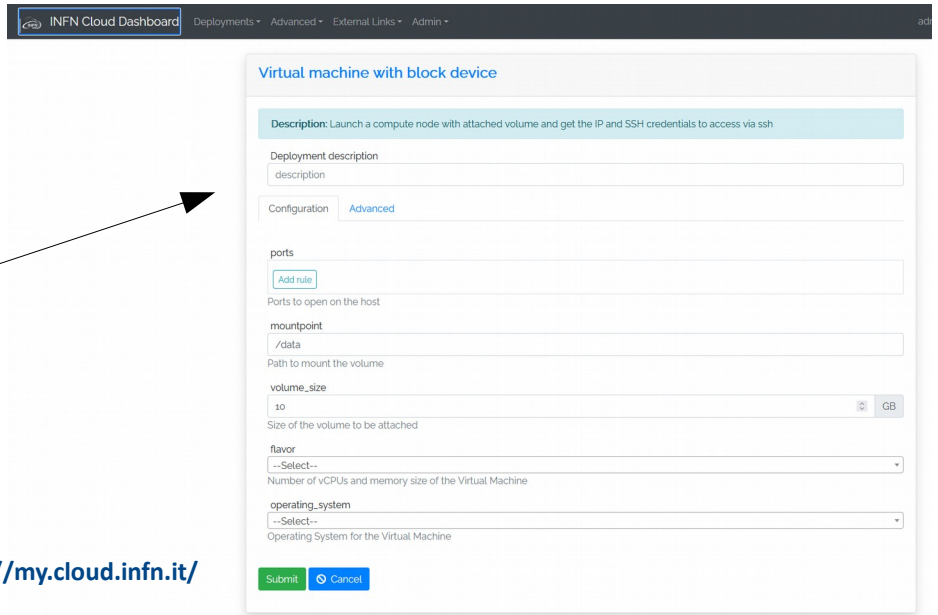
- From a simple VM to the setup of a complex platform configured for experiments.
  - INFN Cloud provides also some centrally managed services such as the Harbor open-source registry, S3 Object Storage and Notebook as a Service.



The dashboard

- Services can be deployed by users via the INFN Cloud easy-to-use web dashboard or via CLI.
- The dashboard hides to the final user all the details about infrastructure and resources allocation complexity. No knowledge about Tosca is required.





Dashboard https://my.cloud.infn.it/

Computing requirements are increasing

- The improvement during these years in performance of CPU and storage versus costs is not going to cover the future computing need for High-Luminosity LHC (2026+), considering a “flat budget” model
  - computing requirements of no HEP experiments are increasing too.
- INFN has to exploit new technical and infrastructural solutions to cope with these requirements → cloud, HPC, data lake: a new model for distributed computing.

To be ready for the next years computing challenges, INFN is revisioning and expanding its infrastructure and services adopting a “cloud first” approach.

  - The target is to create and operate a national vendor-neutral scalable and flexible infrastructure able to serve much more than INFN users and experiments.

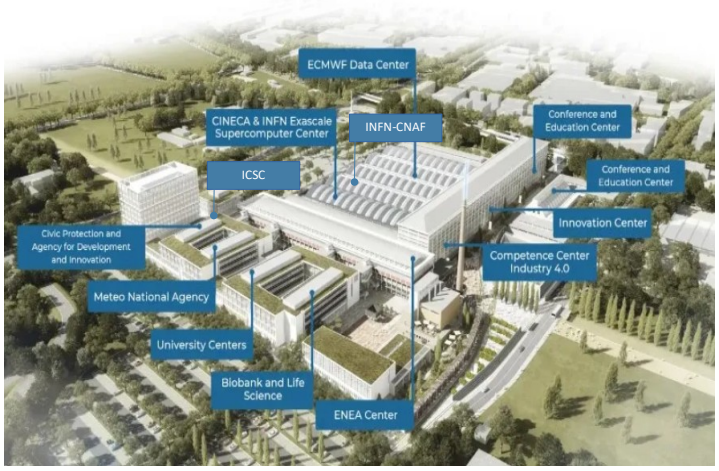
The “Cloud-Data lake” model

- It is the evolution of the computing infrastructure for both Hep and not Hep experiments.
- Resources no longer provisioned only through dedicated grid sites
  - inclusion of HPC systems and commercial clouds as part of the resources we must be able to take advantage of
    - INFN already demonstrated the capability to execute LHC workflows on HPC systems (in particular @ CINECA) and on commercial clouds (e.g. ARUBA).
- Optimize storage access and management
  - reduce the number of replicas, few big sites high speed connected;
  - cpu and storage no longer coupled together;
  - deploy caches where needed.

INFN Cloud is the initial seed of a national data lake infrastructure for research and beyond. It also has a central role in the Italian National Recovery and Resilience Plan (NRRP) computing related initiatives.

Technopole and NRRP: great opportunities

- In the last two years INFN has had the opportunities to promote and participate to projects in the field of infrastructure innovation and technology transfer, with the aim to improve its infrastructure and services.
- Two major opportunities have been come from
  - the realization of the Bologna Technopole
    - it already hosts the Leonardo pre-exascale supercomputer (2022), managed by CINECA-INFN (CINECA Consortium composed by 113 Italian universities and public institutions);
    - it will soon host INFN CNAF and its Tier-1 data center

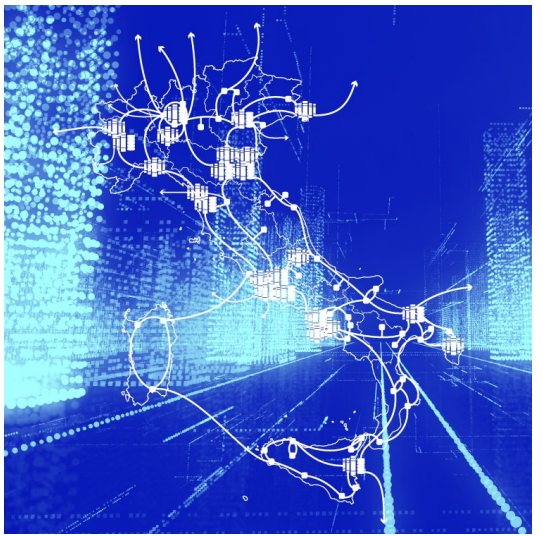


  - the ICSC and TeRABIT projects in the context of the Italian National Recovery and Resilience Plan (NRRP).

Finanziato dall'Unione europea NextGenerationEU
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA

NRRP ICSC National center

- INFN was the proponent of the National Research Center for HPC, Big Data and Quantum Computing (ICSC) set up by the NRRP and financed with €320M.
- The ICSC project is managed through the ICSC Foundation, and includes 51 partners from both public and private sectors.
  - ICSC will be a distributed and traversal supercomputing infrastructure with the mission of supporting the world of scientific research and the business world for the innovation and digitization of the country.
    - Starting from existing HTC, HPC, Big Data infrastructures and evolving to the national Cloud data lake, deriving from INFN Cloud.
  - New resources will be deployed in the existing INFN data centers, and integrated into the national Cloud data lake.



25 Universities

12 Research institutes

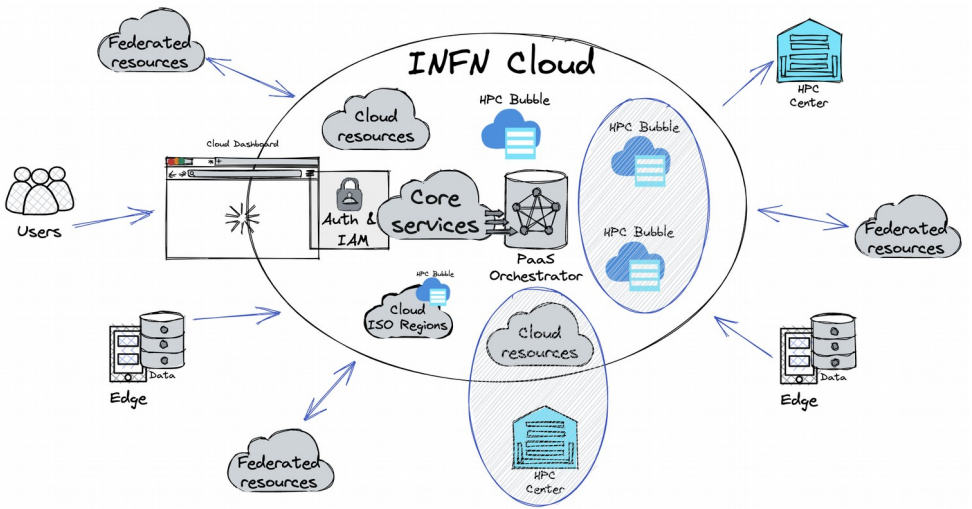
14 Private companies

INFN and Cineca are the leaders of the “infrastructure“ spoke`

NRRP TeRABIT project

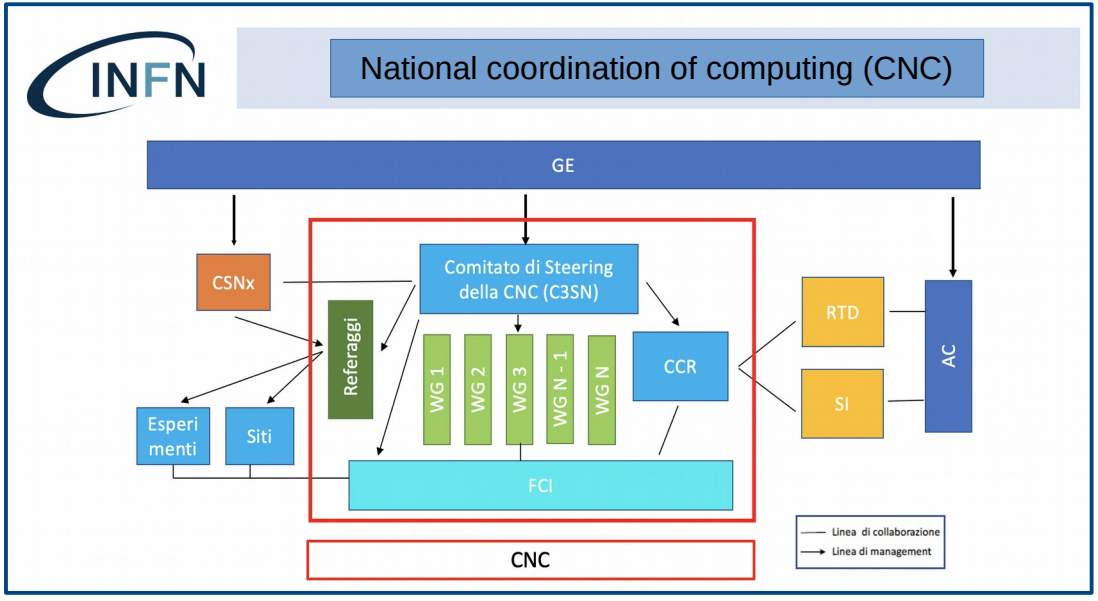
- INFN was the proponent of the “Terabit network for Research and Academic Big data in Italy” project financed with €41M. It aims to create a distributed, hyper-networked, hybrid Cloud-HPC computing environment by integrating the distributed INFN infrastructure with HPC resources of PRACE-Italy (CINECA), through a high speed network provided by GARR and complement the ICSC National Center.
  - INFN will expand its INFN Cloud infrastructure with the deployment of distributed HPC infrastructures ("HPC Bubbles").
    - Clusters with CPUs, CPUs + GPUs, CPUs + FPGAs
    - Fast storage
- Integration between the distributed HPC bubbles;
- integration between the HPC bubbles and the INFN Cloud infrastructure;
- integration between the HPC bubbles and the traditional HPC systems (in particular Leonardo@CINECA).

The aim is to realize a scalable open “Edge-Cloud Continuum”, exploiting AI technologies, that will allow users to process big data in a flexible way.

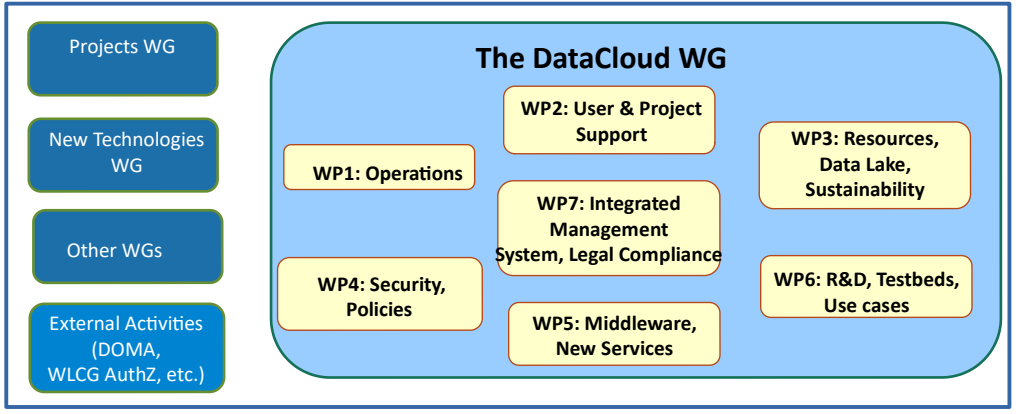


The new INFN computing organization

- INFN decided to reorganize its computing management structure to better cope with these new challenges.



One of these WGs is the "infrastructure" WG, labeled DataCloud, responsible to manage and evolve the INFN distributed infrastructure and services.



DataCloud WG main activities

- Development, implementation and management of the INFN Cloud Data lake architecture.
  - Integration between the traditional WLCG Tiers infrastructure and the “Cloud Native” model represented by INFN Cloud.
  - Support to users and to the management and operation of INFN sites (Grid and Cloud).
  - Development of ISO-Certified solutions.
  - Prototyping, development and support of services starting from use cases.
  - Definition of sustainability models.
  - Cybersecurity: prevention, detection and management of security problems.
  - Taking in account legal and ethical requirements.

Conclusions

- In the next years computing and storage needs of LHC and experiments of other areas will be higher of what existing model can guarantee.
- INFN is actively working on solutions to deal with this challenge adopting the new “cloud – data lake” computing model and exploiting the usage of HPC resources at different scales, possibly also integrating commercial provider resources.
- INFN has the ambition to create and operate a vendor-neutral, open, scalable and flexible “cloud - data lake” that should become the computing environment for fundamental, applied and industrial research in Italy and beyond.

Links

- https://home.infn.it/en/
- https://www.cloud.infn.it/
- https://www.supercomputing-icsc.it/en/icsc-home/
- https://www.terabit-project.it/