ISTITUTO NAZIONALE DI GEOFISICA E VULCANOLOGIA

AI-DAY

9th July 2025

The AI_INFN Platform

ai.cloud.infn.it

AI_INFN
Artificial Intelligence technologies for INFN research

INFN

Lucio Anderlini

Istituto Nazionale di Fisica Nucleare – Sezione di Firenze

INFN Computing Infrastructure, ICSC and TeRABIT

INFN has always relied heavily on computing resources.

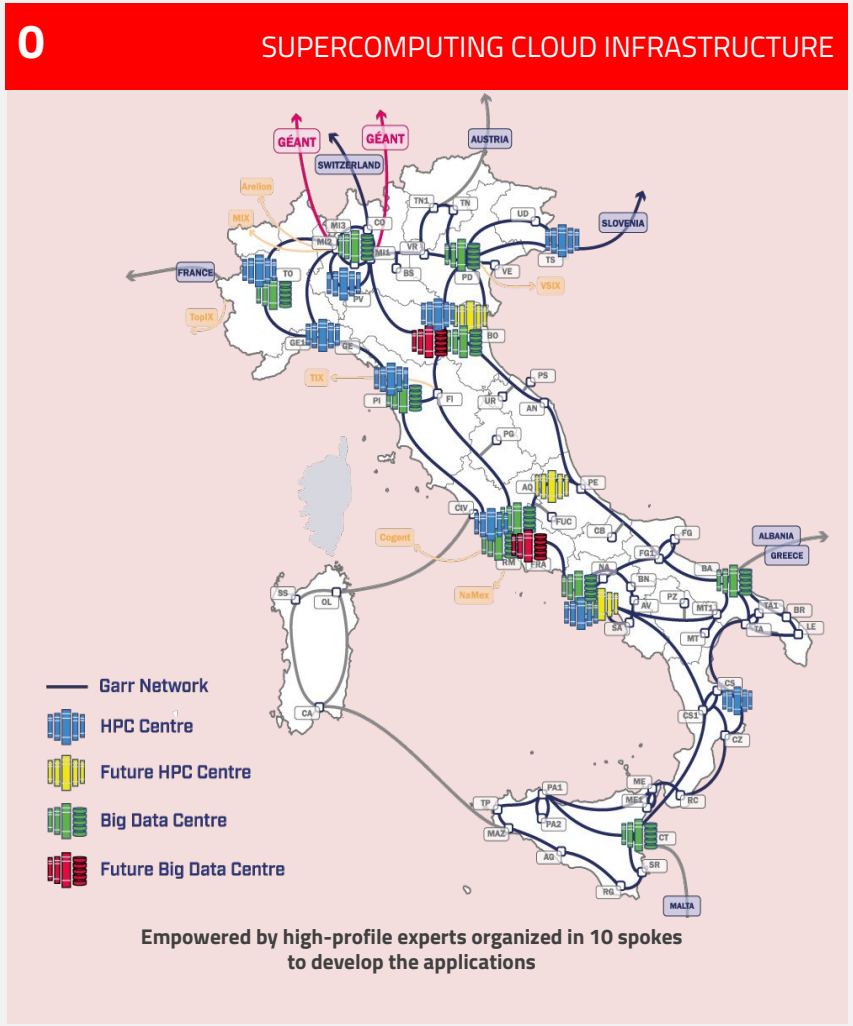
Its infrastructure is distributed by design, fostering collaboration with universities and research institutes.

Boost from NRRP fundings

- **ICSC Supercomputing** – see Arnaud's talk
- **CINECA Leonardo** – Redefining the Italian HPC ecosystem
- **TeRABIT project** – Building a nationally distributed supercomputer through "HPC bubbles."

The pillars of the national computing infrastructure:

CINECA	GARR	INFN
High Performance Computing	*Network*	*Big Data and Cloud Solutions*



INFN

ISTITUTO NAZIONALE DI GEOFISICA E VULCANOLOGIA

AI-DAY

The High Throughput Computing

Historically, INFN has expanded its computing infrastructure in close collaboration with the **Worldwide LHC Computing Grid** (WLCG), a global network supporting data processing for the LHC.

CERN hosts the Tier-0.

INFN CNAF in Bologna hosts Italy's Tier-1 center, a key national hub for data-intensive operations.
Most of the other INFN computing centers are Tier-2.

The LHC experiments produce **hundreds of PB of data per year**

Data is then sent to distributed centers for processing, analysis, and long-term storage.

Operating withing the WLCG fosters:

- Big Data culture
- Data management practices
- Large-scale data processing workflows

Sites of the WLCG

Experiments have dedicated *middlewares* and teams managing the computing operations throughout the WLCG!

ISTITUTO NAZIONALE DI GEOFISICA E VULCANOLOGIA

AI-DAY

INFN DataCloud and the adoption of Machine Learning

In late 2020, INFN introduced **INFN Cloud** as a **private cloud provider** federating sites in *Bologna* and *Bari*.
It complements the WLCG infrastructure with a focus on *smaller experiments* and *individual researchers*.

By 2023, **DataCloud emerged as a flagship model for resource federation** at national level within the NRPP.

DataCloud leverages ***composable***, ***open-source***, ***industrial-grade*** software, simplifying the management of heterogeneous hardware, as GPUs and specialized processors.

→ Immediate appeal for Machine Learning development!

Two specialized initiatives leveraging **DataCloud**:

- 2020 – 2023: ***ML_INFN***, *or* the VM phase
- 2024 – 2026: ***AI_INFN***, *or* the container phase

The mandate of **AI_INFN** is to ease the adoption of Machine Learning techniques for INFN research.

And in particular:

- ✓ Providing **shared**, **low-bureaucracy** access to GPU resources for interactive development;
- ✓ Training (young) researchers with dedicated coding events (*e.g.* **hackathons**)
- ✓ Building a **network of experts** on the adoption of machine learning for science
- ✓ Exploring advanced hardware (*e.g.* **FPGAs** and **Quantum** systems) for ML tasks.

INFN

ISTITUTO NAZIONALE
DI GEOFISICA E VULCANOLOGIA

AI-DAY

INFN DataCloud and the adoption of Machine Learning

In late 2020, INFN introduced **INFN Cloud** as a **private cloud provider** federating sites in *Bologna* and *Bari*.
It complements the WLCG infrastructure with a focus on *smaller experiments* and *individual researchers*.

The mandate of **AI_INFN** is to ease the adoption of Machine Learning techniques for INFN research.

The ultimate goal is to provide
fast, reliable and immediate access to heterogeneous resources,
with a focus on Machine Learning applications.

→ Immediate appeal for Machine Learning development!

Two specialized initiatives leveraging **DataCloud**:

- 2020 – 2023: ***ML_INFN***, *or* the VM phase
- 2024 – 2026: ***AI_INFN***, *or* the container phase

adoption of machine learning for science

✓ Exploring advanced hardware (*e.g.* **FPGAs** and **Quantum** systems) for ML tasks.

INFN

ISTITUTO NAZIONALE DI GEOFISICA E VULCANOLOGIA

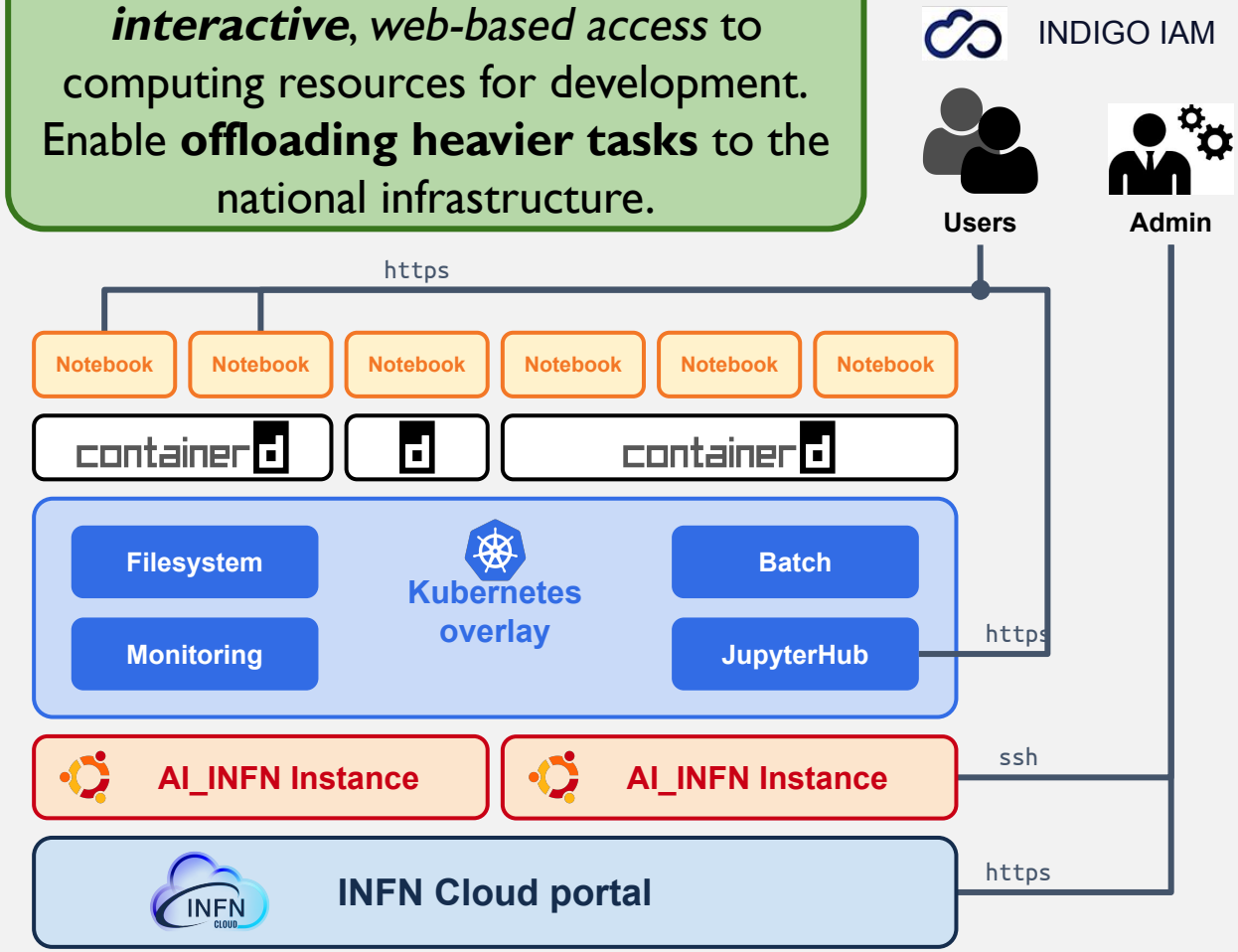
AI-DAY

A cloud-native, GPU-powered, FPGA-equipped platform

The core of the AI_INFN Platform is a dedicated Kubernetes Cluster deployed at INFN CNAF, provisioning hardware accelerators.

CPUs: 180 vCPUs
Memory: 1.2 TB
GPUs: 3x nVidia A100 (up to 21 partitions), 2 nVidia T4 and 5 nVidia RTX 5000
FPGAs: AMD-Xilinx U55c

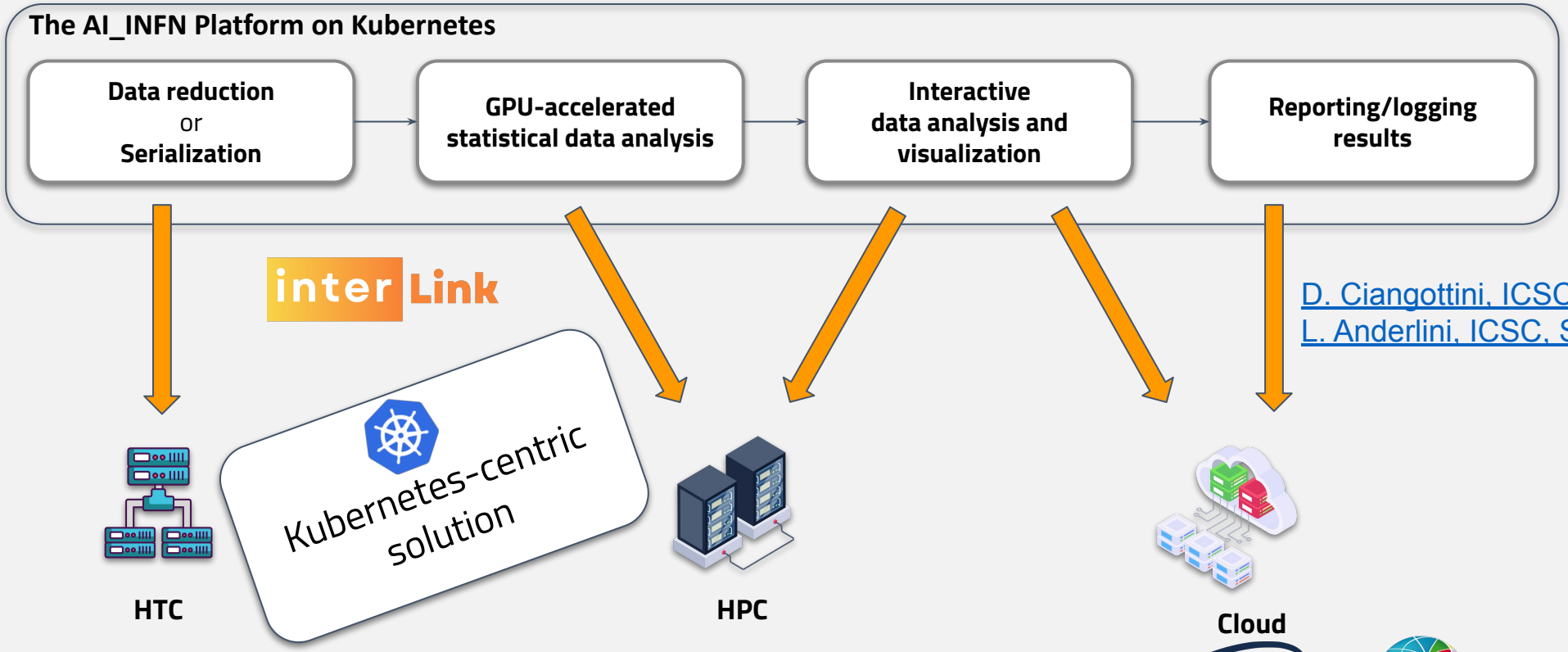
Objective: Provide *immediate*, ***interactive***, *web-based access* to computing resources for development. Enable **offloading heavier tasks** to the national infrastructure.



INFN
ISTITUTO NAZIONALE DI GEOFISICA E VULCANOLOGIA
AI-DAY

Heterogeneity and Distribution: a challenge into an opportunity

The federated computing resources are significantly diverse (HTC, HPC, amd64, arm64, GPUs, FPGAs…)
Assigning tasks to the most suitable resources is key for a cost-effective, large-scale computation.
The AI_INFN Platform offloads tasks to 5 national sites via InterLink: *a plugin-based Virtual Kubelet provider*



D. Ciangottini, ICSC, CHEP 2024
L. Anderlini, ICSC, SIF 2024

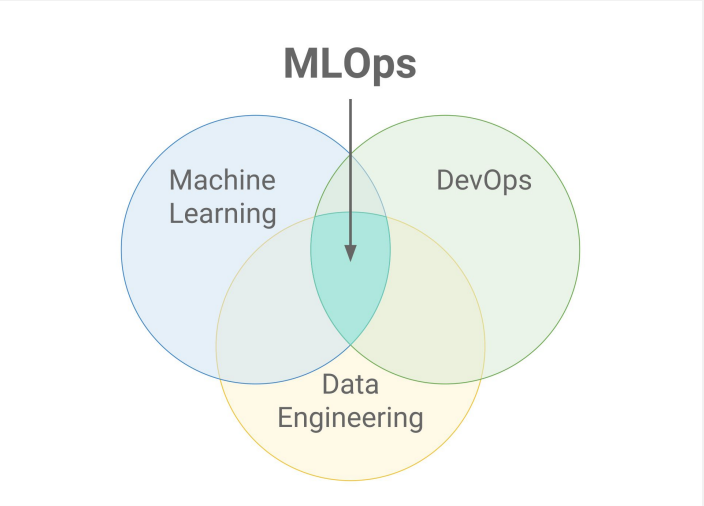
INFN
ISTITUTO NAZIONALE
DI GEOFISICA E VULCANOLOGIA
AI-DAY

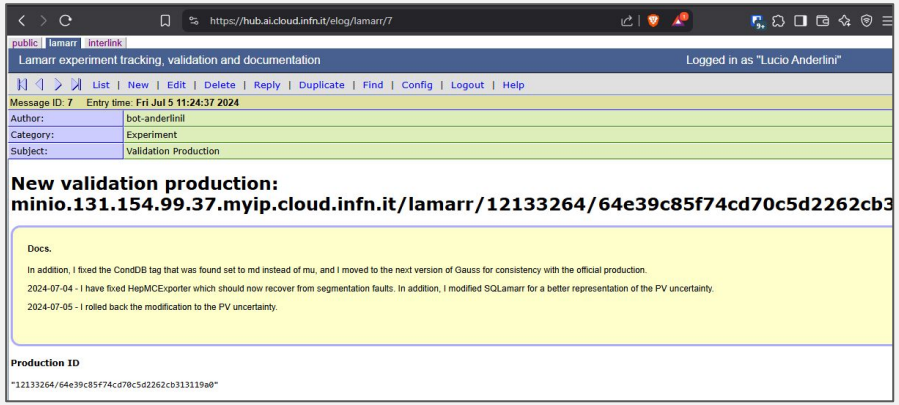
Beyond interactive usage: MLOps

Once developed in the platform, Machine Learning applications should be trivial to distribute.

Most common use cases include:

- DAG management with Snakemake
- Integration with GitHub actions for CI/CD
- Experiment tracking (we are experimenting with PSI ELOG)
- Advanced tooling (all-in-one solutions) such as Kubeflow





snakemake

Kubeflow

GitHub Actions

INFN

ISTITUTO NAZIONALE DI GEOFISICA E VULCANOLOGIA

AI-DAY

Experimenting with data: software and training data

User's homes (software & configuration)

Developing ML requires **interactive tinkering**.

Switching from a GPU model → switch server.

An **responsive network** file system, tuned for intense **metadata activity** is needed:

→ *prototyped* ***bcached Cinder volume via NFSv4***

Data Management for Training & Validation

Training data and metadata are often the most precious asset.

Security + privacy + throughput

Several tools & backends explored:
Ceph+RadosGW (CNAF), Invenio RDM (GLOS), MinIO on K8s (Firenze), xNATS (Pisa), OwnCloud (CH-Net)

Artifact management and distribution

Trained models should be versioned, and distributed.

We prototyped a solution based on **ORAS** (OCI Registry As Storage) for using **INFN Cloud harbor** for artifacts (e.g. trained models).

Automatic upload of artifacts to **cvmfs** via unpacked.

Multi-site distributed file system

POSIX access to data **decouples the AuthN/Z from the application layer**.

Apptainer leverages modern features of the **Linux kernel** to enable mounting custom **fuse volumes**, without admin privileges.

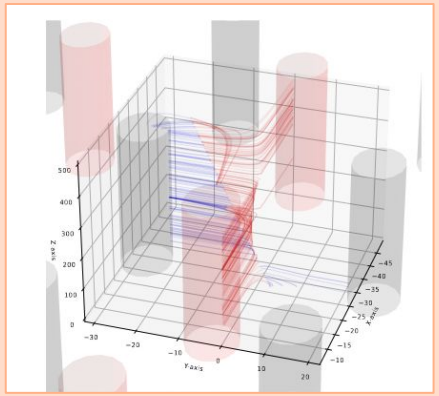
→ Successful integration tests with InterLink.

Selected applications to research topics at INFN

Particle detector modelling – *ICSC Sp. 2 Innovation Grant*

Exploring Physics-Informed Machine Learning to simulate Particle Detectors and Fluid dynamics, relying on GPU acceleration.

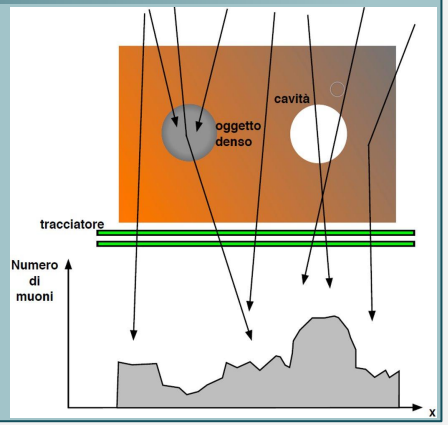
Bombini @ ICSC annual meeting



Muon radiography

Using Convolutional Neural Networks to reconstruct density projections of mines and buildings.

Paccagnella @ IFAE2024



Advanced Hackathon Workshops

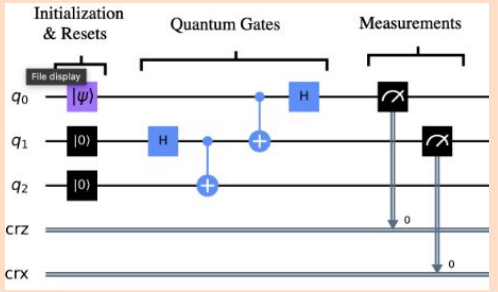
The AI_INFN Platform provide a GPU per participant during the Advanced Hackathons

The latest, in Padova, welcomed 40 students.

GPU-accelerated Simulation of Quantum ML Circuits

The AI_INFN Platform is equipped with several GPU-accelerated simulators of Quantum Circuits used for developing QML.

Cappelli @ CCR 2024



INFN

ISTITUTO NAZIONALE DI GEOFISICA E VULCANOLOGIA

AI-DAY

Conclusion and future work

The AI_INFN Platform provides interactive access to development resources to more than 75 INFN researchers working on more than 20 Machine Learning projects.

It enables **offloading computing payloads** to HTC resources and HPC computing centers, including **CNAF Tier-1**, **CINECA Leonardo** and **TeRABIT HPC bubbles**.

Ongoing effort focused on:

- smooth the submission procedure to remote sites to enhance reliability
- enable community-dedicated deployments of the platform (e.g. small experiments)
- enhanced integration in DataCloud

The experience built with AI_INFN taking part to the **Integration Proof-of-Concept of ICSC Spoke 0** is contributing to the definition of the INFN requirements for future evolution of the national computing landscape (*e.g. the AI Factory, see talk*).