ML_INFN meeting, 18 settembre 2023

The AI_INFN initiative

Lucio Anderlini

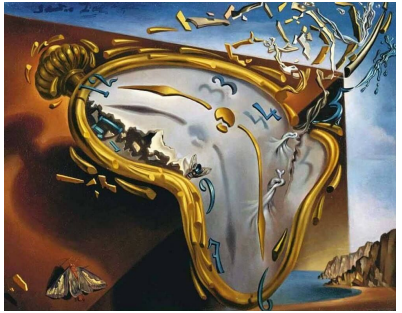
Istituto Nazionale di Fisica Nucleare
SEZIONE DI FIRENZE

State of the art and ML_INFN

The ML_INFN initiative was proposed in 2020 at the dawn of the **INFN Cloud initiative**.

- Commissioned at CNAF a farm with capable of handling **up to 48 simultaneous user** sessions accessing data-center level GPU resources; served via INFN Cloud.
- **Designed and organized 4 educational events** targetting two levels of proficiency (beginner and advanced); highly oriented to discuss the code in small teams.
- Collected and organized examples from success stories of applications of machine learning at research topics in a dedicated web page: **The ML_INFN Knowledge Base**

Four years after, the landscape has changed



- INFN is leading **the ICSC and TeRABIT initiatives**, funded on PNRR resources, exporting the INFN Cloud model to a wider community and wealth of GPU resources, with the name **DataCloud**.
- New models and approaches (*Transformers, Graph Neural Networks, Physics-Informed Neural Networks, Large Language Models, Differentiable Programming…*) have drastically **widened the application range of ML**
- Most Academic Degrees in Physics feature (at least) **entry-level courses on ML** for data analysis, many entry-level courses provided by *Ufficio Formazione*
- New hardware and computing technologies are arising as "*specialized accelerators*" for performing machine learning at scale: ***Quantum Computing* and *FPGAs***.

It's time to renew ML_INFN to make it ready for the upcoming challenges!

WP1 Infrastructure and Resource Provisioning

Lots of resources coming from ICSC and TeRABIT?

→ Less "*pressure for being in production*" on our farm

→ Opportunity for contributing to the provisioning model

Focus shifts towards R&D on the provisioning model, with a systemic view to ease ML workloads.

Needs for an updated and well maintained farm.

Open Science and Advanced Education WP2

What will be added-value in our *hackathons*?

→ ML_INFN has attracted a community of world-leading experts in the application of ML to research in physics

→ We canore ambitious in the target of our *hackathons,* letting experts to discuss their code

Focus shift towards *Advanced Hackathon Workshops.*

AI_INFN
Artificial Intelligence technologies for INFN research

Scientific use cases

Applications to scientific research remains central.

To develop the tools for making it easier to do machine learning for INFN researchers,
we need them to use to tools and provide feedback.

WP3 *How?* User support and community engagement

ML on FPGA and Quantum Computers

New hardware will change the landscape of computing.

Deploying ML algorithms of **FPGAs** enables fixed-latency, low-energy inference.

Quantum Computing will enable extremely fast computations of specialized, possibly trained, algorithms.

WP4

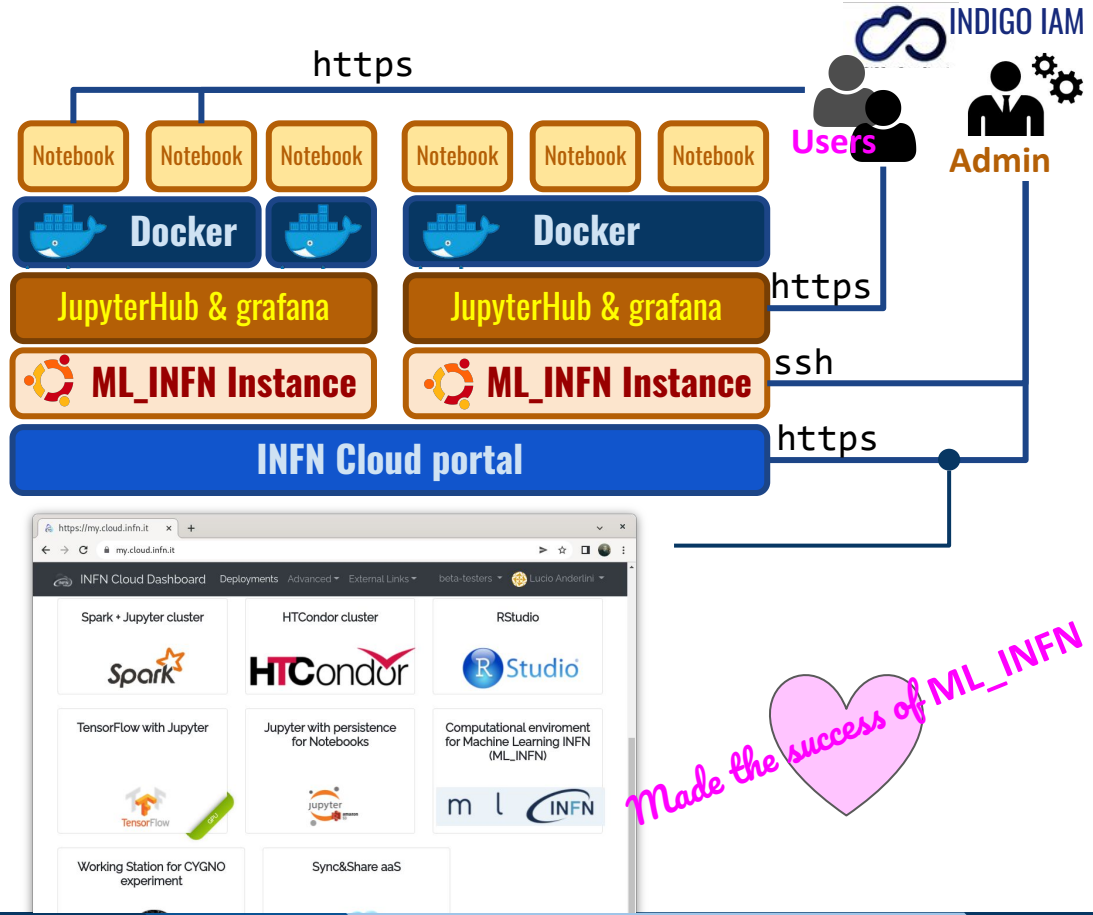
The provisioning model: ML_INFN version

Each project gets its own Virtual Machine

At the end of the project, the VM is destroyed, the GPU is freed for other users/projects, data in the filesystem is lost.

Resources are **guaranteed to the project**

Inefficient and **too many admins.**



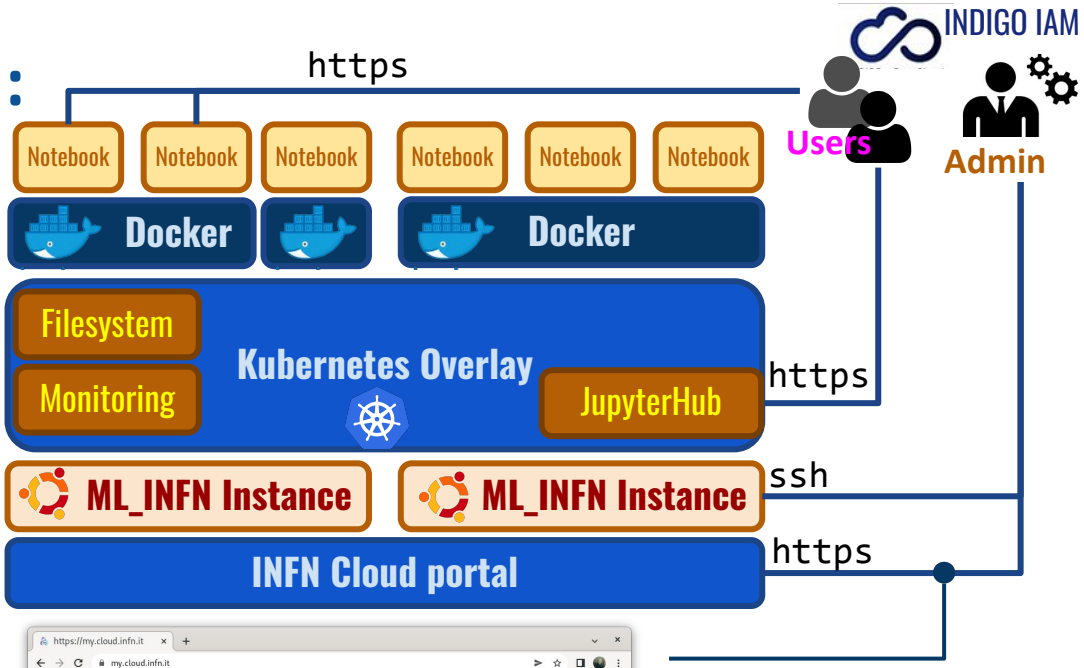
The provisioning model: AI_INFN version

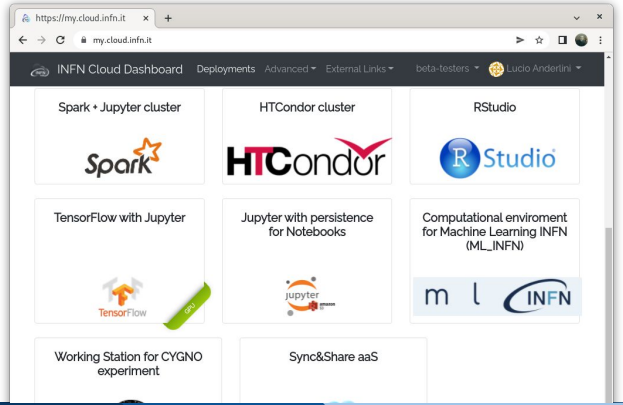
An additional abstract, elastic overlay is added on top of multiple VMs.

Adding and removing machines enables **scaling based on demand**.

Filesystem is persistent at platform-level: GPUs can be re-assigned without data loss.

Guarantee access to resources will require custom policies.





Sinergy with CLOUD_ML project (ECRF)

(some) Scientific use cases

Machine-Learning based Ultra-Fast Simulation for the LHCb experiment

Advanced models inspired from automatic translation models to simulate the response of LHCb Calorimeter, accounting for particle-to-particle correlation effects

Sinergy with LHCb (CSN1)

Model-independent searches for New Physics with Domain Adaptaion

Deep classifiers to distinguish signal from background with an efficiency made explicitly independent of the coupling of a new Higgs-like particle.

Sinergy with CMS (CSN1)

Image recoloring from XRF scans

XRF scans are processed with Convolutional Neural Networks for inferring the visible color of damaged or covered painture labels

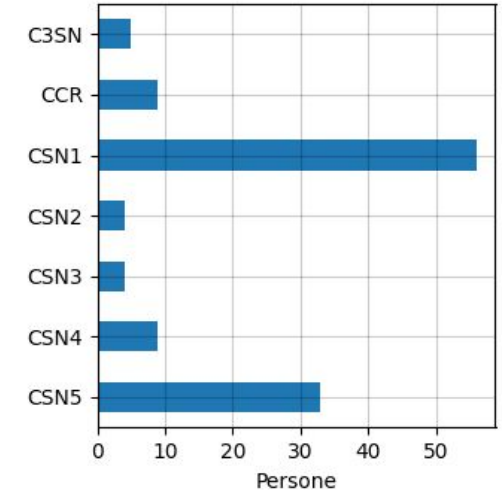
Sinergy with CHNet and LABEC

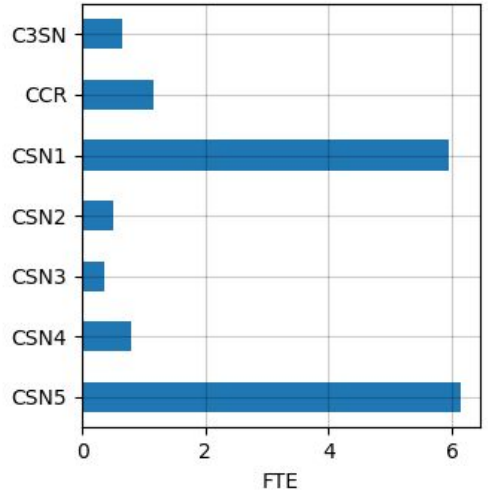
Anagrafica

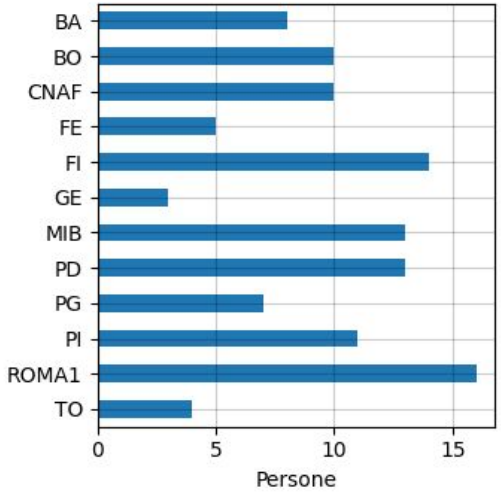
80 ricercatori e 42 tecnologi
(+ sinergie importanti con ICSC, TeRABIT e FAIR)

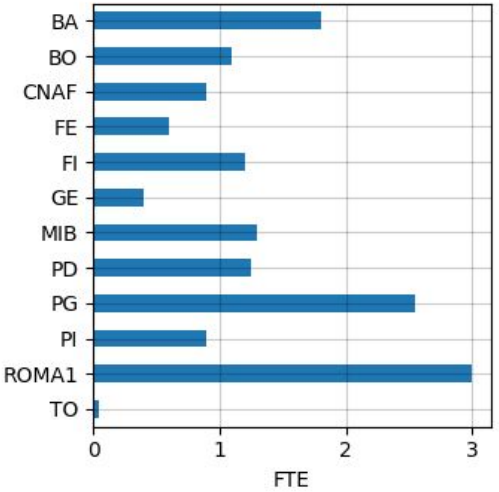
Unità coinvolte e Resp. Locali

- BA - Alfonso Monaco
- BO - Daniele Bonacorsi
- CNAF - Stefano Dal Pra
- FE - Enrico Calore
- FI - Lucio Anderlini (Responsabile Nazionale)
- GE - Luca Rei
- MIB - Simone Gennai
- NA - Francesco Alessandro Conventi
- PD - Marco Verlato
- PG - Daniele Spiga
- PI - Francesca Lizzi
- ROMA1 - Stefano Giagu









Richieste finanziarie

Aggiornamento e manutenzione della farm:

- Nuovo server: 30 k€ / anno al CNAF
- 2x FPGA Alveo v70

Missioni per *Advanced Hackathon Workshop*:

- 1 k€ / Struttura / anno
- + 4 k€ su Firenze

Nota più “burocratica”:

Su suggerimento di Alessandra Retico, abbiamo inserito la richiesta per il nuovo server in CALC5_TIER1.

La richiesta però è un po’ ibrida e non è stato chiarissimo dove collocarla.

CALC5_TIER1 normalmente raccoglie richieste di risorse informatiche acquisite indipendentemente dal CNAF, che possano essere messe a disposizione dell’esperimento.

In questo caso, stiamo chiedendo un server specifico ***compatibile con macchine pre-esistenti*** *secondo requisiti tecnici ben definiti, che pure andrà collocato al CNAF e che potenzialmente potrà offrire anche risorse ad altri, ma che si collochi nella* ***medesima*** **tenancy** ***dei server di CSN5 acquistati nel 2020****.*

Valutazioni preliminari (da CSN5): *Impatto scientifico/tecnico.*

La continuazione delle attività di condivisione delle expertise relative all'AI entro l'INFN è di sicuro interesse.

Per migliorare l'impatto scientifico delle attività proposte è fondamentale promuovere i seminari in maniera capillare.

Si suggerisce di mantenere dei corsi entry-level all'interno degli anni del progetto.

CHNet Medea

Il progetto Medea, incentrato su tecniche di **NLP per beni culturali**, non ha incontrato i favori della CSN5.

Su suggerimento di CSN5, la RN E. Ronchieri (CNAF) sta valutando l'ingresso in AI_INFN

La modalità è in via di definizione, propendiamo per una formalizzazione del concetto di *"use-case"* in WP3, con un responsabile ben definito. **Medea sarebbe uno *"use-case"*.**

La parte "beni culturali" fatica a trovare collocazione in AI_INFN, mentre la parte **"NLP" è di sicuro interesse per altre attività INFN** (ad esempio analisi dei log del Tier1).

Superset strutture a cui estendere la partecipazione ad AI_INFN: **RM3, LNGS, LNS**

Richieste finanziarie (**fino a 3 k€/anno di missioni + 1 k€/anno per servizi OpenAI**)

Backup

Deliverables

D1.1 Overlay Kubernetes distribuito su più macchine virtuali;

D1.2 Infrastruttura di monitoring e accounting interna a AI_INFN;

D1.3 Batch system per uso opportunistico delle risorse non coinvolte in attività di sviluppo;

D1.4 Prototipo integrabile tra i servizi gestiti di *DataCloud*.

D2.1 Organizzazione di un *Advanced Hackathon Workshop* di apertura;

D2.2 Sviluppo di un corso base da fruire in modalità *e-learning*;

D2.3 Organizzazione di un *Advanced Hackathon Workshop* di aggiornamento;

D2.4 Organizzazione di un *Advanced Hackathon Workshop* di chiusura.

D3.1 Organizzazione di seminari periodici su applicazioni di *Machine Learning* ai temi di ricerca di rilevanza per l'Ente;

D3.2 Identificazione di una soluzione di supporto-utenti per l'utilizzo della piattaforma;

D3.3 Implementazione della soluzione di supporto-utenti identificata;

D3.4 Report di valutazione consuntiva sul modello di provisioning sviluppato.

Milestone	
M1.1	31 / 12 / 2024
M1.2	30 / 06 / 2025
M1.3	31 / 12 / 2025
M1.4	31 / 12 / 2026
M2.1	31 / 12 / 2024
M2.2	31 / 12 / 2024
M2.3	31 / 12 / 2025
M2.4	31 / 12 / 2026
M3.1	31 / 12 / 2026
M3.2	30 / 06 / 2024
M3.3	31 / 12 / 2024
M3.4	31 / 12 / 2026

Deliverables (WP4)

D4.1 Dimostratore operativo di acceleratori FPGA fruiti tramite Cloud;

D4.2 Sviluppo e documentazione nella *Knowledge Base*, di tecniche di compressione e ottimizzazione (occupazione risorse FPGA, latenza e throughput nella fase di inferenza) di modelli classici di *Machine Learning* e *Deep Learning* per utilizzo su acceleratori FPGA commerciali;

D4.3 Esempio di *Quantum Machine Learning* documentato nella *Knowledge Base*;

D4.4 Sviluppo di metodologie basate su *Machine Learning* classico per la preparazione, ottimizzazione (*e.g. transpiling*, simulazione realistica di sorgenti di errore), e *quantum error correction*, di circuiti quantistici di tipo NISQ, documentate nella *Knowledge Base*;

D4.5 Dimostratore operativo di interfaccia tra INFN Cloud e le risorse di quantum computing da fornitori esterni (per esempio IBM, o risorse ICSC).

M4.1	30 / 06 / 2025
M4.2	31 / 12 / 2024
M4.3	31 / 12 / 2025
M4.4	31 / 12 / 2026
M4.5	31 / 12 / 2026