Developing an automated ATLAS analysis workflow on the INFN Cloud facility

Meeting federazione ATLAS-IT

26.VII.2023

Leonardo Carminati, Caterina Marcon, David Rebatto, Ruggero Turra

INFN MILANO
Istituto Nazionale di Fisica Nucleare
Sezione di Milano

Motivation

- ATLAS has been using a complex and distributed computing infrastructure: the **Worldwide LHC Computing Grid (WLCG)** characterized by almost a million computing cores and an exabyte of storage deployed in different sites worldwide;

- The computing needs (power and storage) of ATLAS in the **HL-LHC** era will represent an **unprecedented challenge** for the existing infrastructure:

  - New software and hardware technologies are being explored;
  - The experiment is considering integrating various alternative computing resources into the distributed computing system, including **cloud computing technologies**.

- Cloud technology allows **dynamic, flexible and cost-effective resource provisioning**.

Motivation

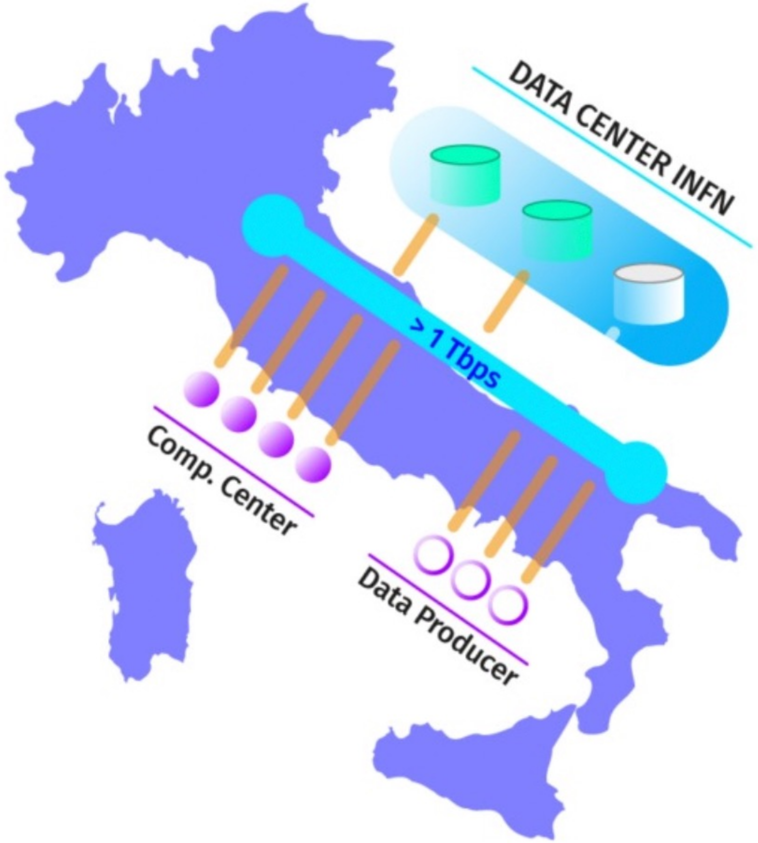
- INFN Cloud is the initial seed of a **National Data Lake** and it has a central role in the computing initiatives planned within the scope of the **National Recovery and Resilience Plan (PNRR)**;

- With the funds of the **ICSC** and **TeRABIT** projects, INFN will renew its distributed infrastructure (consisting of 1 first level centre, 9 second level centres and 2 specific target centres) by incorporating these resources into a **unique cloud ecosystem**;

- As a long term goal, this layout will allow transparent and efficient access to resources for all users, using a **PaaS or IaaS approach**.

INFN Cloud infrastructure

- INFN CLOUD infrastructure in production since March 2021;
- Backbone connecting the large data centres of CNAF and Bari;
- Smaller federated sites offer opportunistic resources;
- Resources orchestrated by OpenStack;
- Active INFN users can access all the federated resources;
- Appointed “administrators” can provide sub-services;
- Two operation models:
    - Platform-as-a-Service (PaaS)
    - Software-as-a-Service (SaaS)



INCANT: INfn Cloud based Atlas aNalysis faciliTy

Objectives and Tools

- Investigate the possibility of implementing two distinct (but not orthogonal) analysis workflows by exploiting the computational resources of **INFN Cloud**:
  - create a batch-like system capable of obtaining flat n-tuples (compatible with analysis flows for result extraction) from structured and complex data;
  - develop **interactive analysis flows** (similar to Jupyter Notebook-as-a-Service).

- High level building blocks:
  - Different **Docker images** to create an alternate **ATLAS software stack provisioning** architecture;
  - Using **Kubernetes** for **resource orchestration**;
  - Using **HTCondor** as the **job scheduling system**.

The Platform-as-a-Service paradigm

Users maintain Applications and Data

+

Providers maintain the underlying infrastructure

=

PaaS paradigm

- Applications
- Data

User

- Runtime
- Middleware

INFN CLOUD + R&D project

- OS
- Virtualization
- Servers
- Storage
- Networking

INFN CLOUD

CERN - ATLAS middleware and runtime added on top of existing PaaS services from INFN Cloud

The R&D resource pool

- The following resource pool has been provisioned:

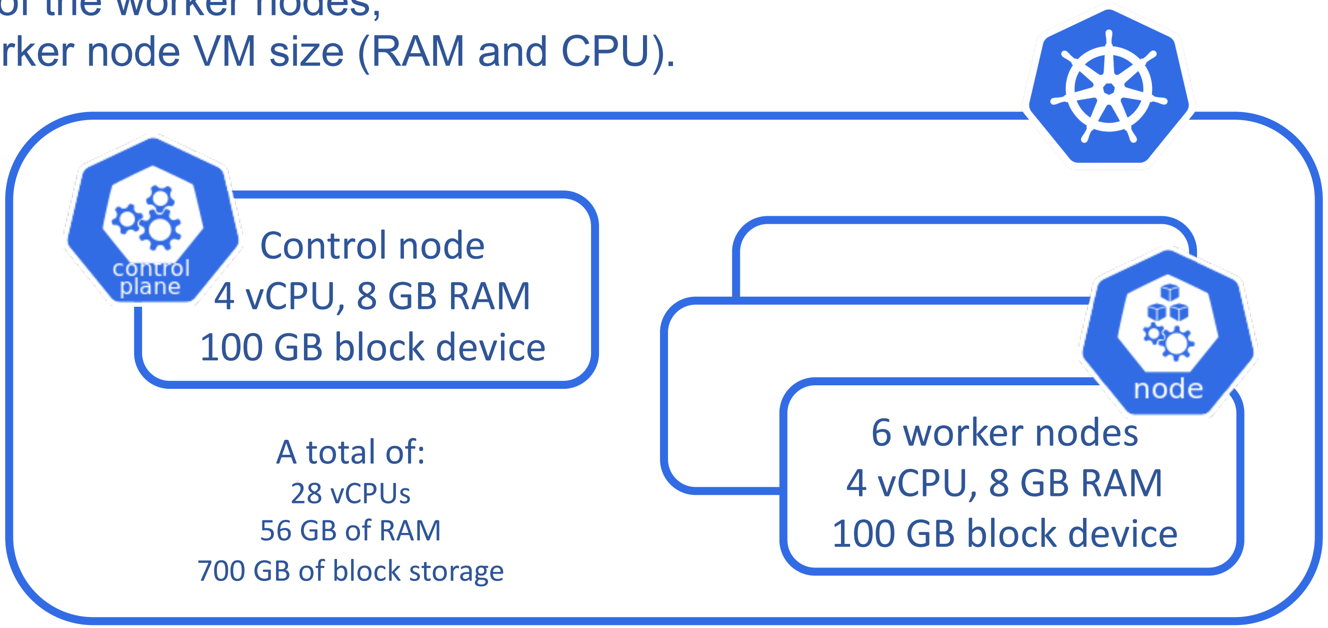
CPU	92
RAM [GB]	168
Volumes [GB]	1000
External storage (compatible with S3) [GB]	2048

- A pre-defined set of cloud applications is available:
  - pure Kubernetes clusters;
  - HTCondor clusters deployed on Kubernetes;
  - General purpose Virtual Machines (with Ubuntu 18.04, Ubuntu 20.04 or CentOS 7);
  - S3 storage.

- The scale of these applications is configurable and resources are drawn from the reserved pool.

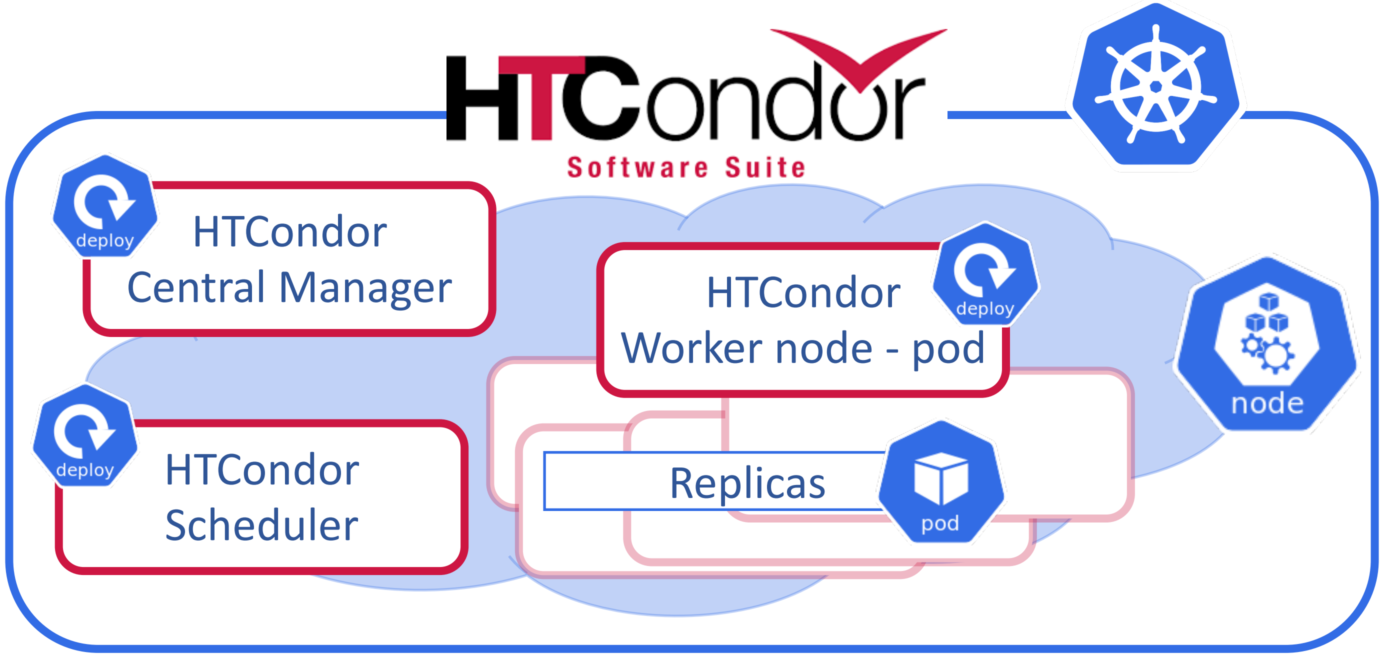
HTCondor on Kubernetes (I)

- **Kubernetes** (K8s) cluster with **HTCondor batch system** on top created via INFN Cloud Dashboard;
- Resources drawn from R&D pool and orchestrated by OpenStack
- Basic monitoring services configured by default (e.g. Grafana dashboard with Prometheus);
- Limited user configurability:
  - number of worker nodes;
  - Docker image of the worker nodes;
  - master and worker node VM size (RAM and CPU).



HTCondor on Kubernetes (II)

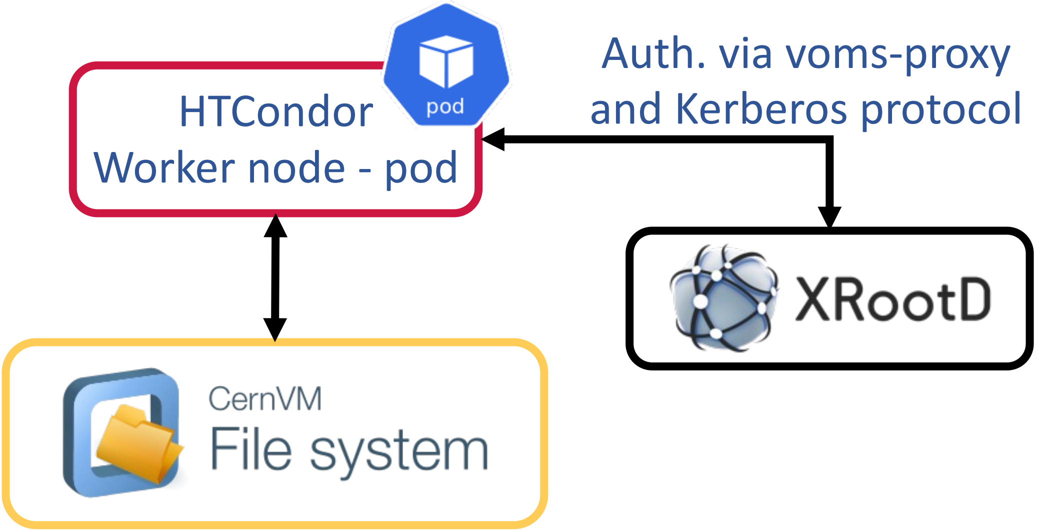
- HTCondor components configured as **K8s deployments**;
- Deployments can be easily **scaled** by the cluster administrator;
- **No HTCondor submit node** on cluster by design to allow remote job submission.



HTCondor on Kubernetes (III)

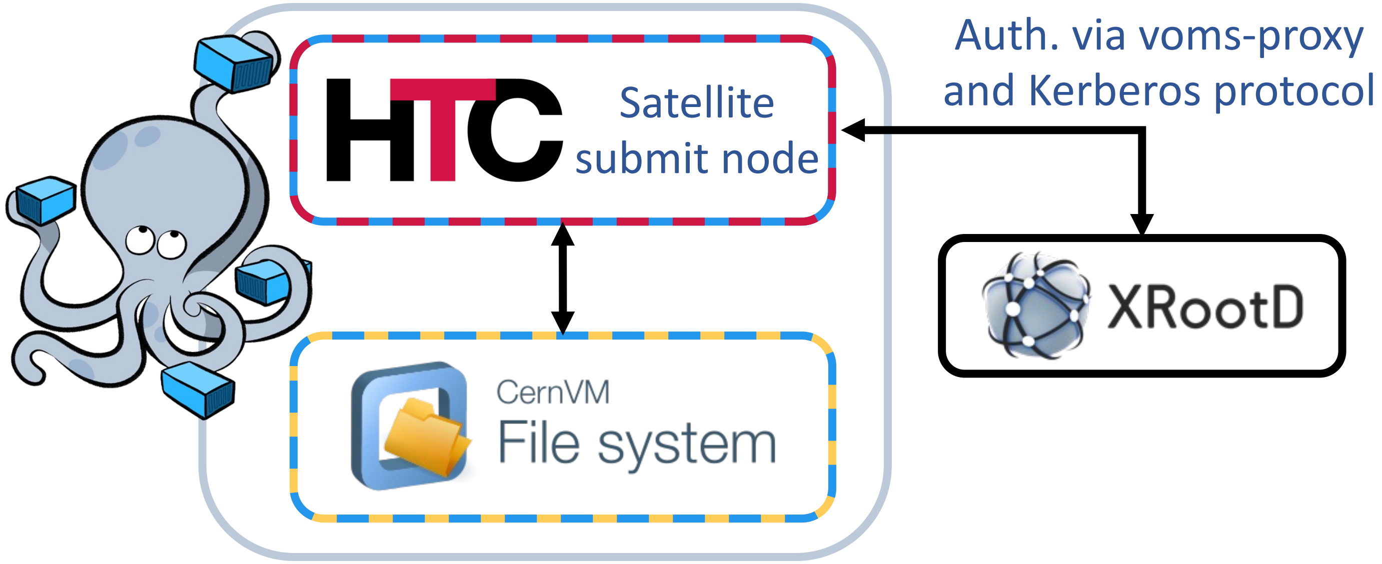
- Submit node designed as a **satellite Docker container**;
- Jobs can be submitted to the cluster from any **remote location**;
- Authentication to the HTCondor cluster via the **INFN IAM infrastructure**.

Merging CERN and INFN resources (I)



- ATLAS resources must be linked:
  - CVMFS to retrieve the required software;
  - XRootD for data file transfer.
- HTCondor worker pod images updated to include CVMFS and support for X509 and Kerberos authentication.

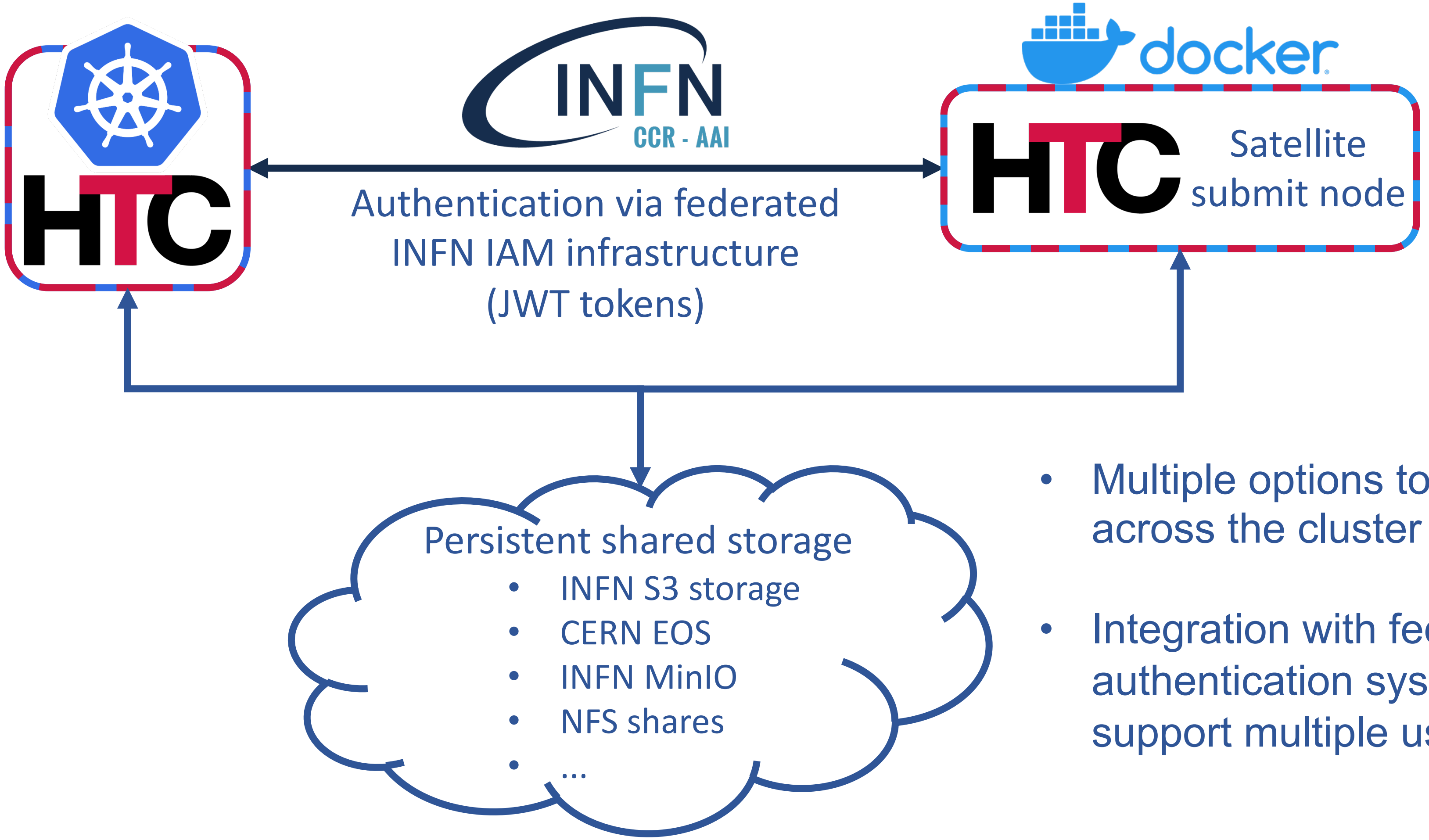
Merging CERN and INFN resources (II)



- CERN authentication for XRoot access added to the submit node;
- CVMFS running in a separate container using the official cvmfs/service: 2.10.1-1 image [1], [2];
- Container integration via Docker Compose;
- Host machine and containerized ecosystem are isolated (except for the shared kernel).

[1] https://hub.docker.com/layers/cvmfs/service/2.10.1-1/images/sha256-511c85a96c50f89871dbfc1ebd9ab1d7df6b54310cc3745b9043e08fbfabea89
[2] https://cernvm.cern.ch/fs/

External storage integration



- Multiple options to share data across the cluster nodes;
- Integration with federated/SSO authentication systems crucial to support multiple users.

Eureka!

- The following configuration has been tested and promising results have been obtained:
  - a VM with Ubuntu 20.04, the *storage controller*, has been configured as a NFS server;
  - the storage controller is also equipped with RClone, which is connected to a 2 TB, S3-compatible bucket, hosted on INFN resources. The S3 storage system also provides a web interface;
  - all cluster nodes, as well as the remote submit node, can natively mount the NFS share as a POSIX drive by means of standard Linux libraries;
  - the functionality of the HTCondor cluster when operating without a shared filesystem is also being evaluated.

Future steps

- Define a strategy to **extend the availability** to a larger user base;
- Start a **thorough testing phase**:
  - Computing performance;
  - Scalability;
  - Behaviour under different type of analysis workloads;
  - Flexibility to support different workflows.
- Create a **pre-configured environment** easily deployable on INFN Cloud;
- Add **interactive workflow capabilities** to the batch-like system:
  - Batch system for data pre-processing and simplification;
  - Interactive platform for data manipulation and result extraction (plots, charts, etc.).

Conclusions

- The increasing need for computing resources foreseen for the HL-LHC era is accelerating the development of new analysis strategies based on distributed computing systems, including **cloud computing technologies**;

- In 2021, INFN joined this effort and the new **INFN CLOUD infrastructure** started operations;

- The core of the infrastructure is fully managed by INFN and users/administrators have access to **PaaS** and **SaaS** solutions, such as K8s/HTC clusters, VM or Jupyter Notebook-as-a-Service applications;

- A new R&D project has started to extend the existing PaaS paradigm with **ATLAS data analysis capabilities**;

- The objective of this project is to support **batch-like and interactive analysis workflows**, ensuring an optimized use of computing resources;

- So far, a working prototype has been produced and **testing is ongoing**.

Acknowledgments

This R&D project is funded by INFN Milano, INFN CLOUD and supported by ATLAS Italia Calcolo.

caterina.marcon@mi.infn.it

External storage integration

- Several constraints must be considered:
  - **purpose**: a shared storage intended for long-term archiving has different requirements from a low-latency drive used for intense computing;
  - **ease of access and configuration**: nodes running on the Kubernetes cluster don't allow interactive access and must not require constant supervision. Therefore authentication workflows must be self-sustaining and capable of handling token refresh automatically;
  - **possibility of remote access**: the submit node is outside the Kubernetes/HTCondor fence but must nevertheless be able to access the shared resources seamlessly. This entails that it is necessary to secure the connection of the submit node to the INFN CLOUD resources.