Developing an automated ATLAS analysis workflow on the INFN Cloud facility

Mini-Workshop ATLAS Italia Calcolo

28 – 29 Novembre, 2023

INFN Genova

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;

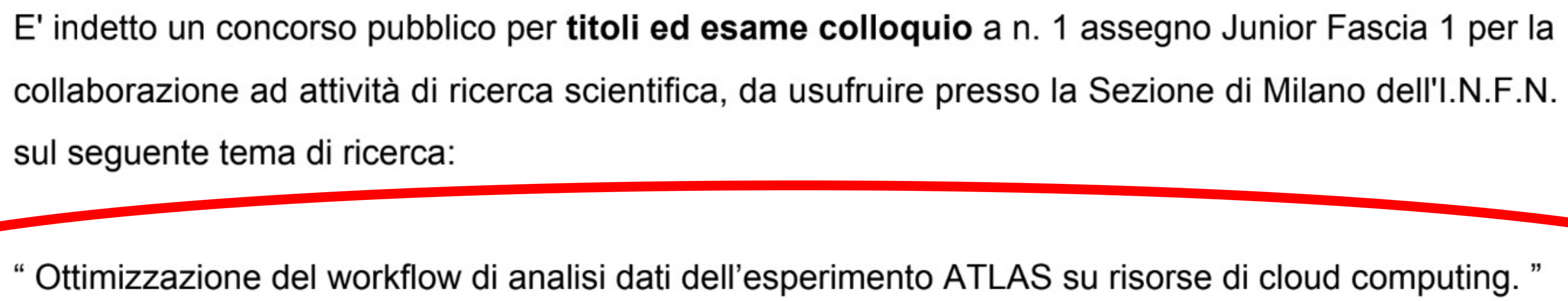
- The computing needs (power and storage) of ATLAS in the **HL-LHC** era will represent an **unprecedented challenge** for the existing infrastructure:

  - 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**.

Where did the project come from?

- In March 2021, a research grant was awarded with the aim of **developing and optimizing analysis workflows of the ATLAS experiment on cloud computing resources**:



E' indetto un concorso pubblico per **titoli ed esame colloquio** a n. 1 assegno Junior Fascia 1 per la collaborazione ad attività di ricerca scientifica, da usufruire presso la Sezione di Milano dell'I.N.F.N. sul seguente tema di ricerca:

" Ottimizzazione del workflow di analisi dati dell'esperimento ATLAS su risorse di cloud computing. "

- Although not explicitly specified in the job description, it came "natural" to think of exploiting the resources made available by **INFN CLOUD**.

Where we were…

- One year ago…

Individuare gli aspetti basilari per lo sviluppo di un'unità di base

Sviluppo di un'unità di base

Individuare altri aspetti caratterizzanti

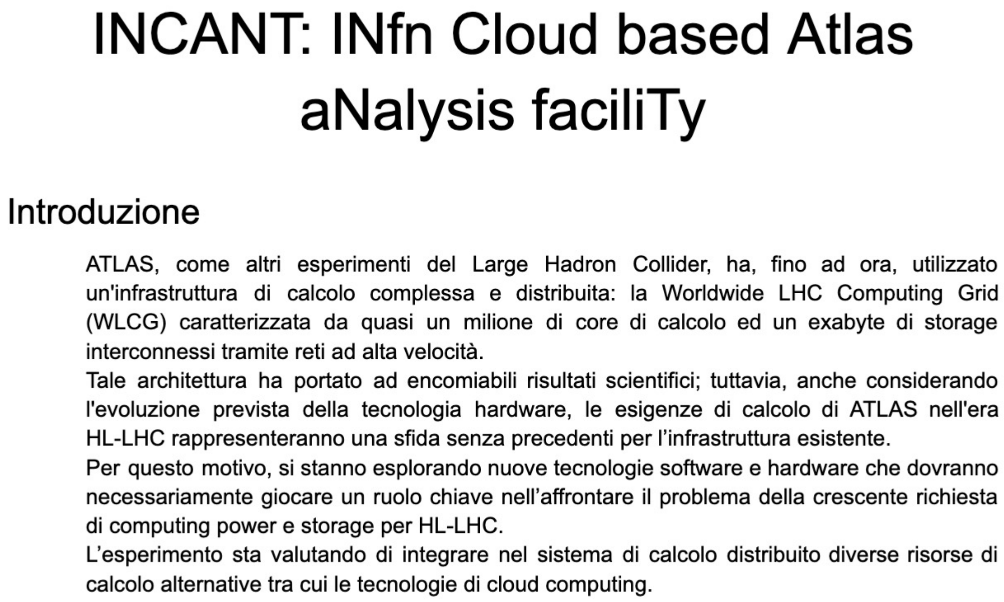
Prototipo di prodotto funzionante

Prodotto finito + Utenti

We were here…

INCANT: INfn Cloud based Atlas aNalysis faciliTy

- February 2022 the INCANT Project is presented to the INFN Cloud Board [1];
- March 2022 the INFN Cloud Board has approved the Project and provided the requested resources for 3 (+3) months.



INCANT: INfn Cloud based Atlas aNalysis faciliTy

Introduzione

ATLAS, come altri esperimenti del Large Hadron Collider, ha, fino ad ora, utilizzato un'infrastruttura di calcolo complessa e distribuita: la Worldwide LHC Computing Grid (WLCG) caratterizzata da quasi un milione di core di calcolo ed un exabyte di storage interconnessi tramite reti ad alta velocità.
Tale architettura ha portato ad encomiabili risultati scientifici; tuttavia, anche considerando l'evoluzione prevista della tecnologia hardware, le esigenze di calcolo di ATLAS nell'era HL-LHC rappresenteranno una sfida senza precedenti per l'infrastruttura esistente.
Per questo motivo, si stanno esplorando nuove tecnologie software e hardware che dovranno necessariamente giocare un ruolo chiave nell'affrontare il problema della crescente richiesta di computing power e storage per HL-LHC.
L'esperimento sta valutando di integrare nel sistema di calcolo distribuito diverse risorse di calcolo alternative tra cui le tecnologie di cloud computing.

[1] https://docs.google.com/document/d/1jjXWMFfR9y7tT0wFc7OLkpYzQt9yBEfIj-pET3KVmGQ/edit#heading=h.9xaw9ljbkncm

INCANT: INfn **C**loud based **A**tlas a**N**alysis facili**T**y

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 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.

The Platform-as-a-Service paradigm

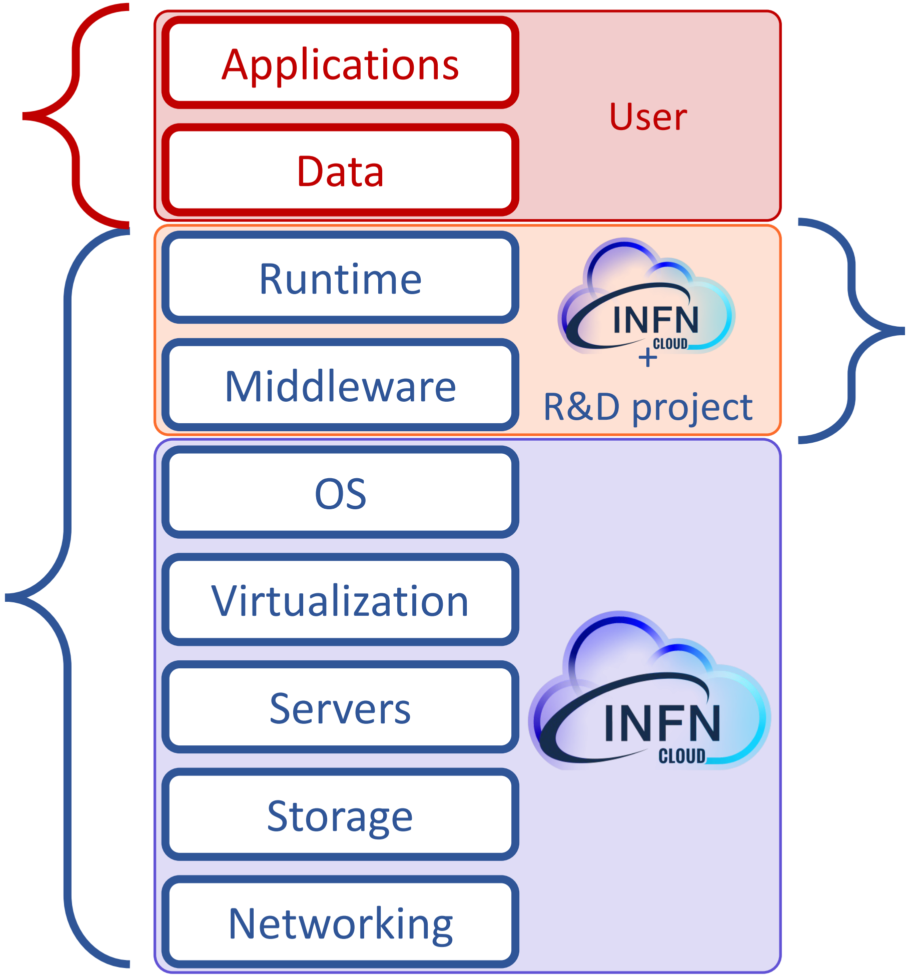
Users maintain Applications and Data

+

Providers maintain the underlying infrastructure

=

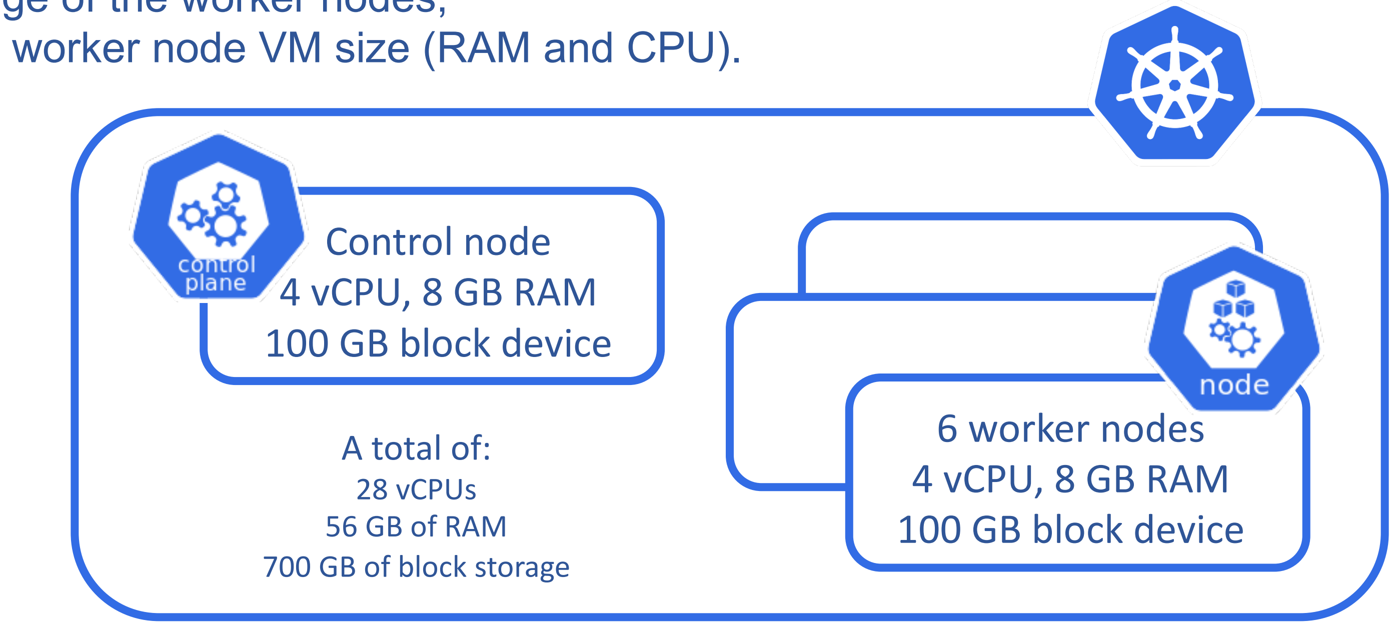
PaaS paradigm



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

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).



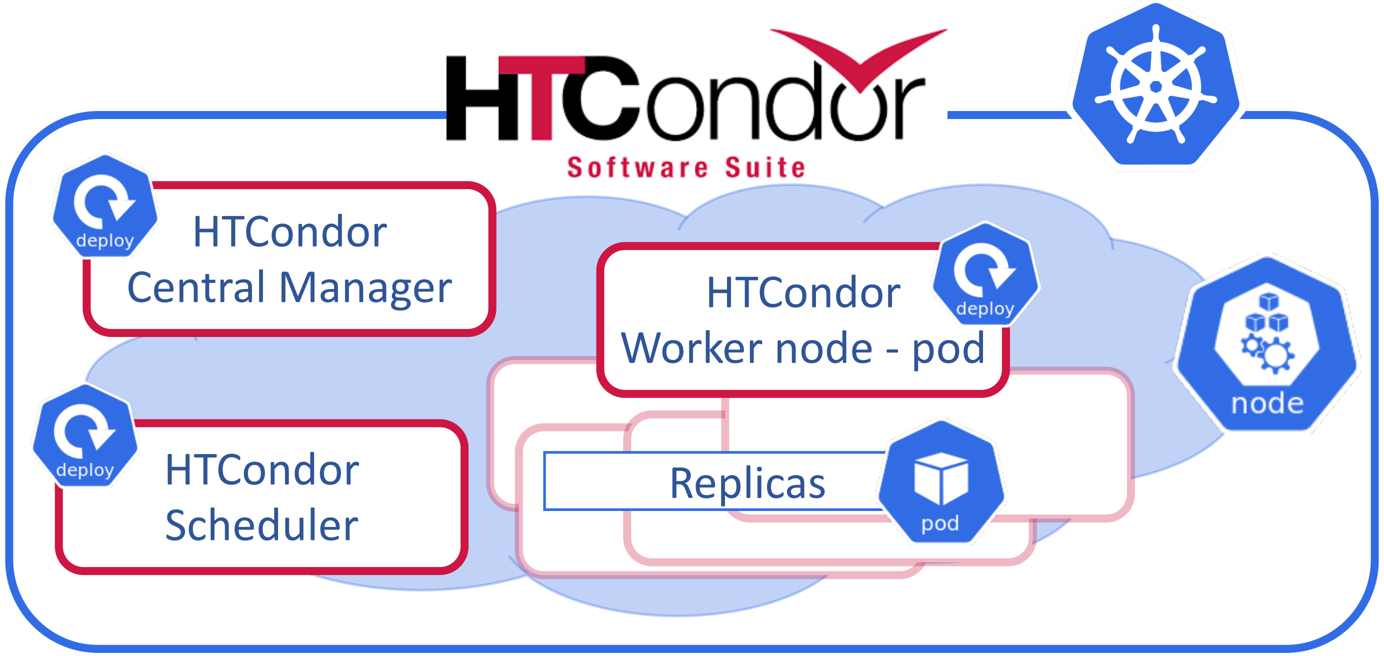
Control node
4 vCPU, 8 GB RAM
100 GB block device

A total of:
28 vCPUs
56 GB of RAM
700 GB of block storage

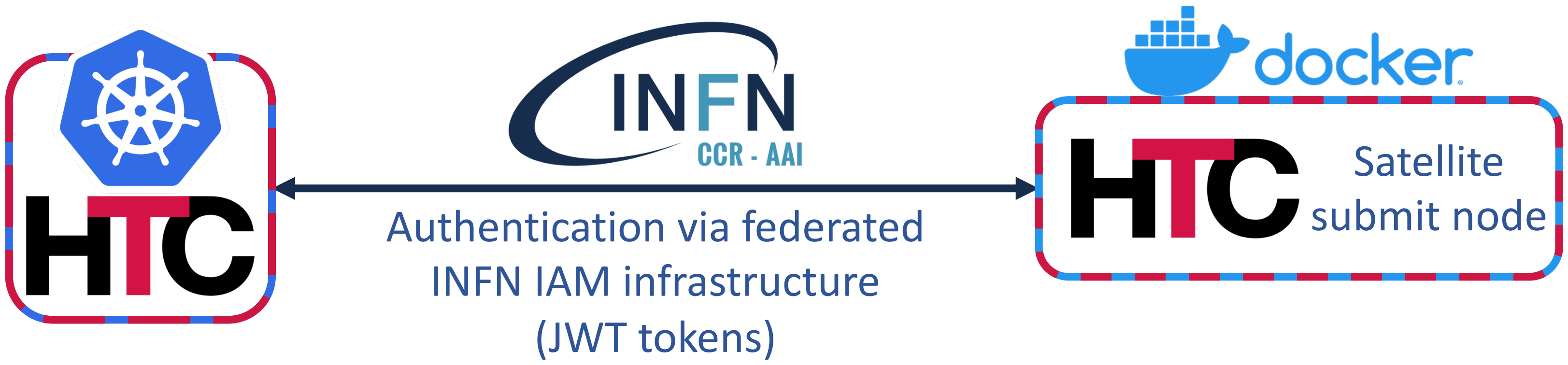
6 worker nodes
4 vCPU, 8 GB RAM
100 GB block device

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.



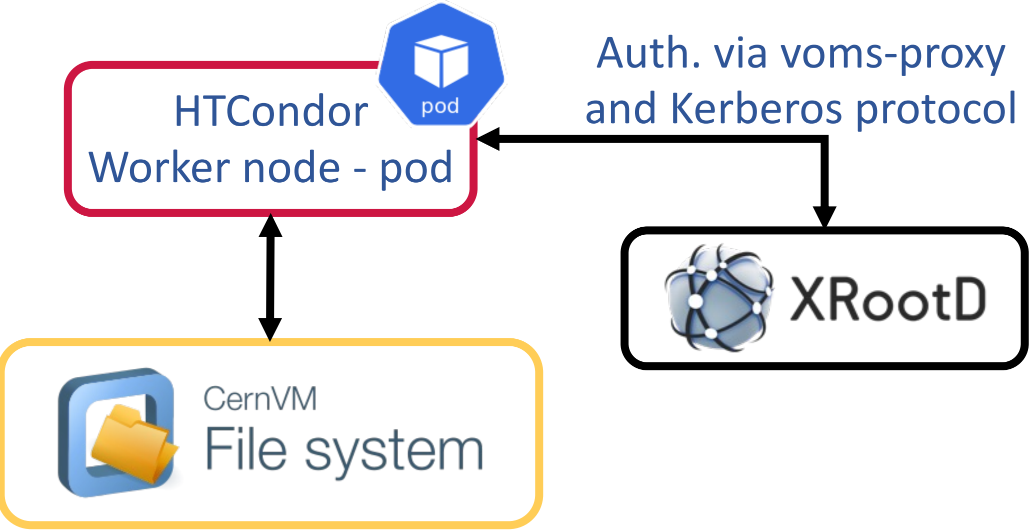
HTCondor on Kubernetes (III)



Authentication via federated INFN IAM infrastructure (JWT tokens)

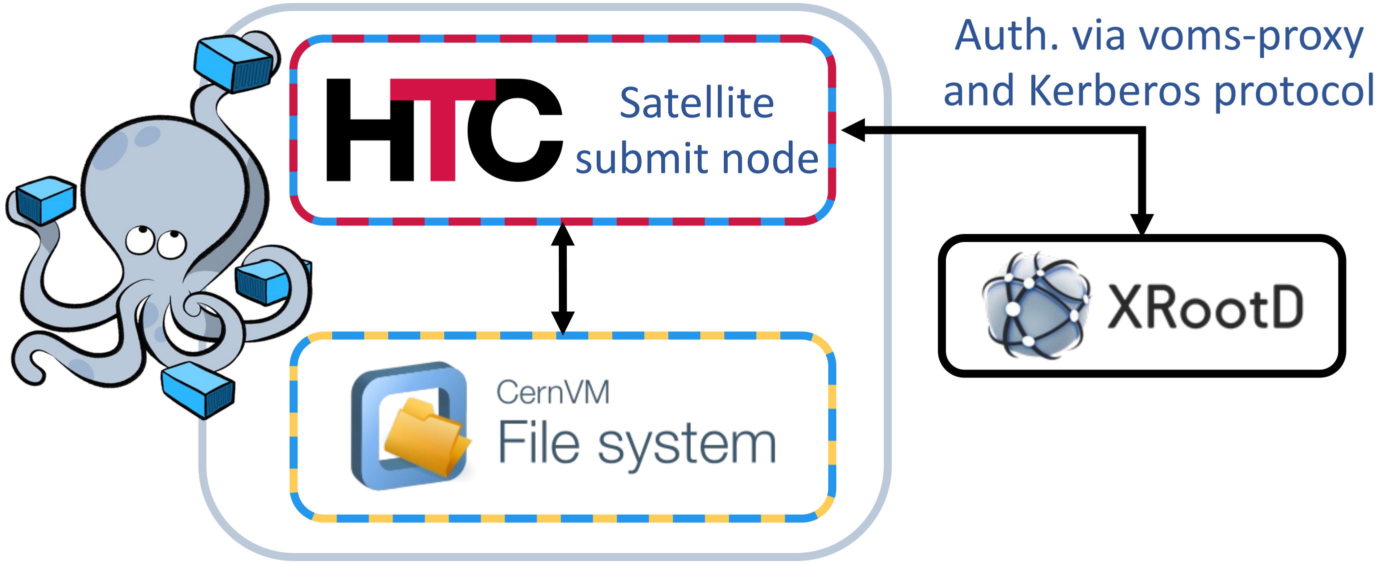
- Submit node designed as a **satellite Docker container**;
- Jobs can be submitted to the cluster from any **remote location**;
- Authentication to the HTCondor cluster via the **INFN IAM infrastructure**.

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.

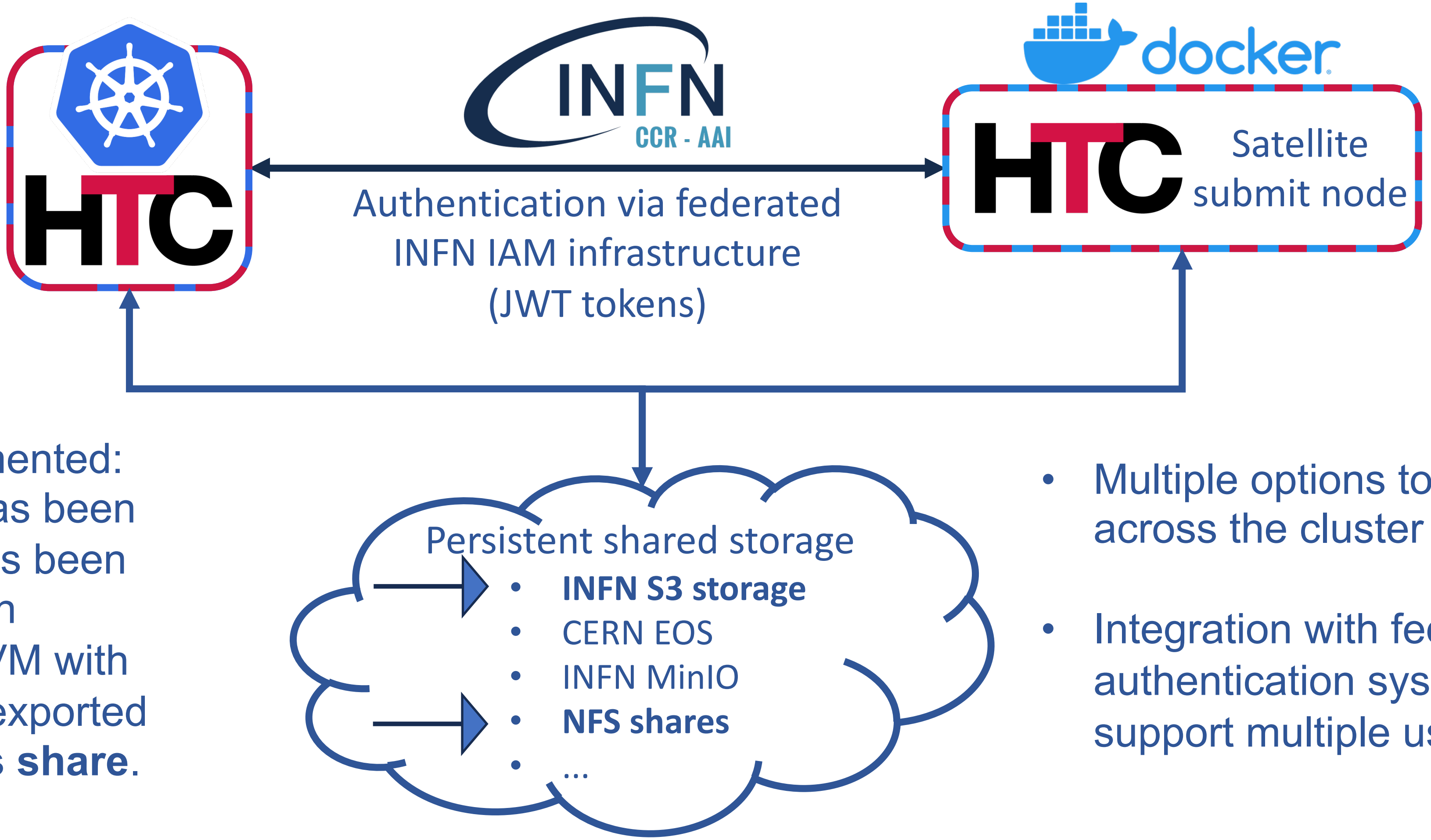
Merging CERN and INFN resources (II)



- 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



- Solution Implemented:
  S3 storage has been provided; it has been mounted on an intermediate VM with **R-clone** and exported again as a **nfs share**.

- Multiple options to share data across the cluster nodes;

- Integration with federated/SSO authentication systems crucial to support multiple users.

Outcomes

- On the INCANT resources, a **typical ATLAS workflow has been tested:**
  - Data read from EOS has been processed using the **common CERN framework** (cvmfs, asetup, athena, etc) and saved the output locally.

- This project also allowed find and fix an HTCondor ATLAS Driver bug (https://its.cern.ch/jira/browse/ATLASG-2560);

- Development of INCANT results in ~25 tickets with INFN Cloud Service Desk;

- The project has been presented in many different occasions:
  - Oral presentation at CHEP23: https://indico.jlab.org/event/459/contributions/11508/
  - Peer reviewed proceeding (already accepted): https://indico.jlab.org/event/459/contributions/11508/
  - Workshop CCR Loano: https://agenda.infn.it/event/34683/abstracts/23448/
  - ATLAS ITALIA (poster) https://web.infn.it/atlas/

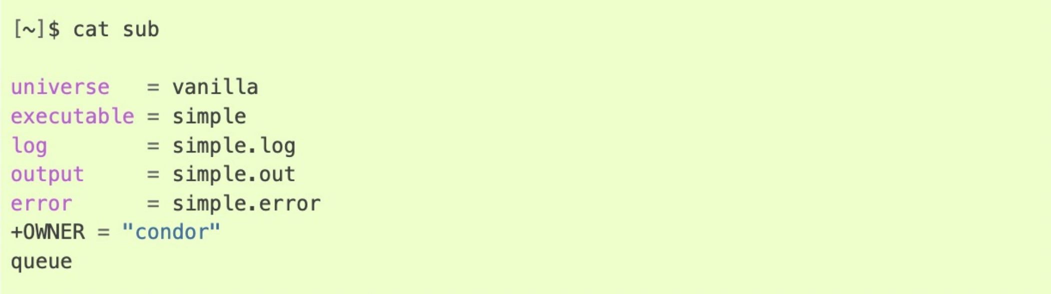
Critical issues found in the infrastructure

- Catch-all partition:
  - The creation of VM instances was not always straightforward;
  - Some instances on the ReCas - Bari, after updates requested by the system, were no longer reachable (neither from outside nor from inside);
  - The creation of ad-hoc clusters starting from the instance of single VMs creates some critical issues: if the machines are instantiated on different sites, the public IPs of the machines on one site are not visible to the machines on the other site.

Critical issues found in the infrastructure

- INCANT Partition:

1. Obsolete versions (e.g. the HTCondor version);
   - It is also difficult to update the various versions because the K8s + HTCondor cluster tends not to restart after carrying out the updates;
2. Multi-user management:
   - All containers run with administrator privilege;
   - As the project was conceived, HTCondor must be able to manage INFN users who have affiliation with the ATLAS experiment. However, it is specifically indicated to set the "owner" option for the HTCondor job submission [1];



```
[~]$ cat sub

universe   = vanilla
executable = simple
log        = simple.log
output     = simple.out
error      = simple.error
+OWNER = "condor"
queue
```

   - It is necessary to implement the possibility of segregating the jobs of each user: by implementing a permission/access system that guarantees data confidentiality.

[1] https://guides.cloud.infn.it/docs/users-guides/en/latest/users_guides/sysadmin/compute/htcondor.html#job-submission

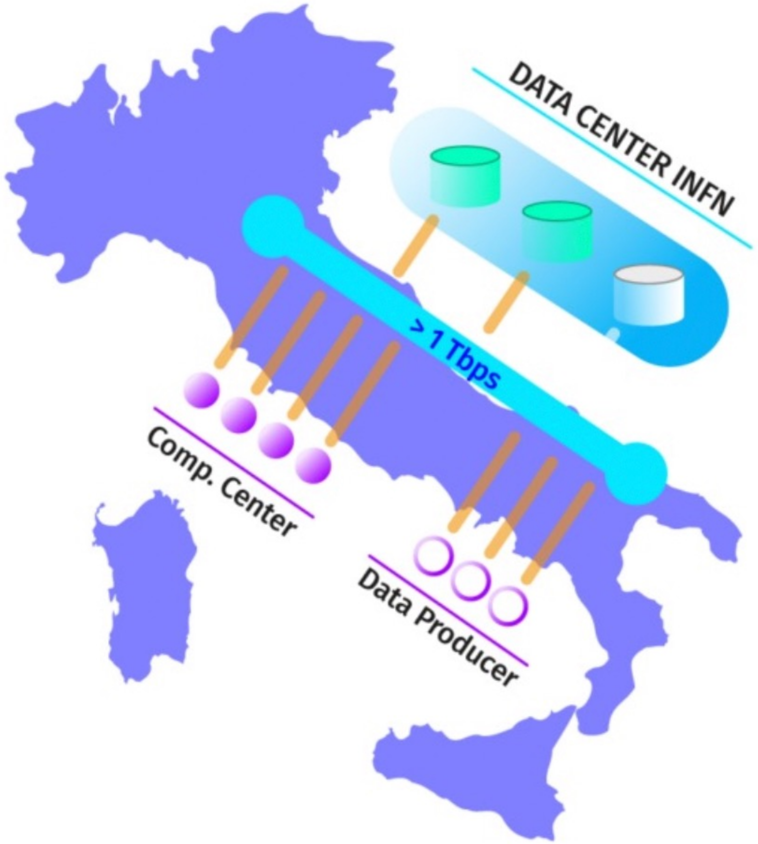
Next steps (?)

- Le risorse INCANT sono state assegnate dal Board di INFN Cloud per 3 (+3) mesi…ora stanno per scadere ☹

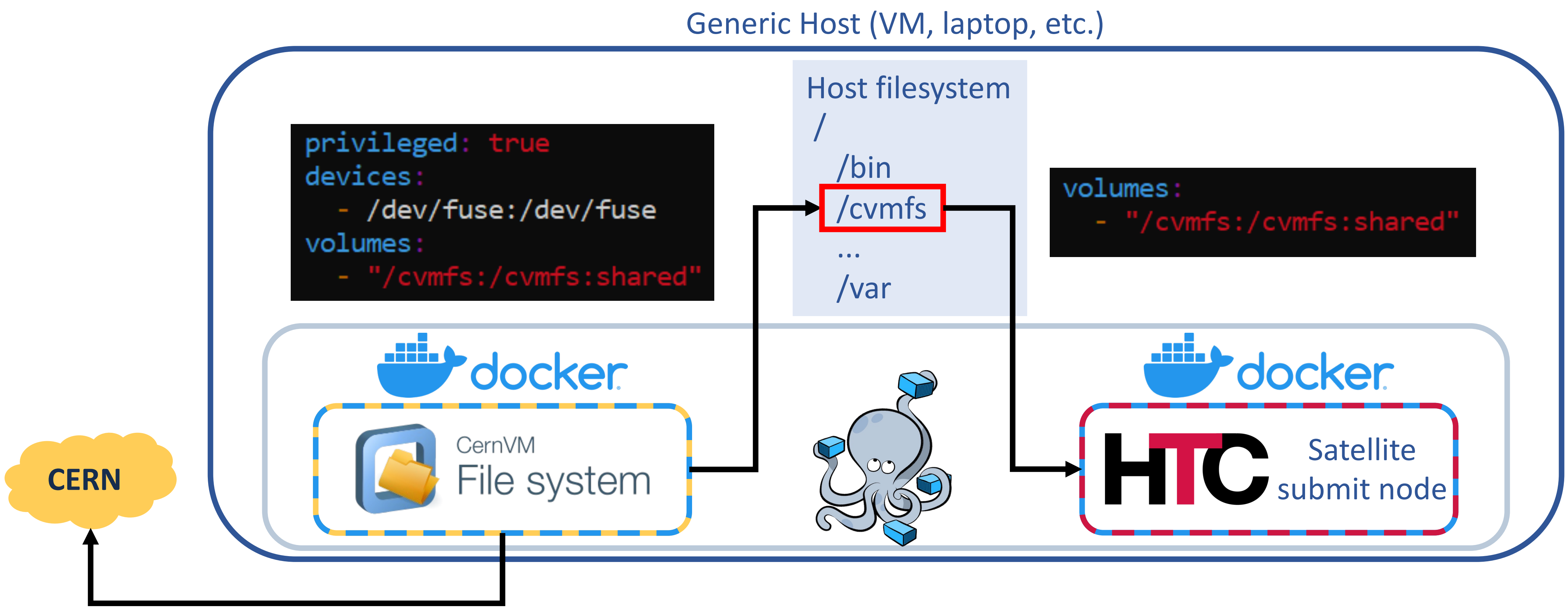
> *Il PMB ritiene che sia necessario che* ***il progetto INCANT faccia richiesta ufficiale per le risorse cloud alla CSN1 o al C3SN.***
> *Infine, alla luce di esigenze tecniche dell'infrastruttura Cloud@CNAF, il PMB* ***chiede di distruggere tutte le VM INCANT entro il 15 dicembre 2023,*** *che si potrebbe quindi anche configurare come data di fine concessione.*
> *Quindi disattiveremmo il progetto il 15/12 e provvederemmo a ri-attivarlo dopo che avrà superato il referaggio.*

INFN Cloud infrastructure

- INFN CLOUD infrastructure in production since March 2021;
- Backbone connecting the large data centers 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)



Merging CERN and INFN resources



Generic Host (VM, laptop, etc.)

```
privileged: true
devices:
  - /dev/fuse:/dev/fuse
volumes:
  - "/cvmfs:/cvmfs:shared"
```

```
Host filesystem
/
  /bin
  /cvmfs
  …
  /var
```

```
volumes:
  - "/cvmfs:/cvmfs:shared"
```