INFN k8s activities

A. Ceccanti, D. Ciangottini, C. Duma, D. Spiga
on behalf of INFN-PG, CNAF, INFN-Cloud teams

Outline

This presentation is organized in different sections that follow the schema below:

- K8s usage for compute/data integration in data-lake: *the DODAS model*
  - Data processing facilities:
    - Compute
    - Storage and caches
  - The example of CMS experiment and other early adopters
- K8s for long running services
  - Deploying IAM on K8s
  - STORM@CNAF on K8s
- K8s @ INFN-Cloud
- Outlook

K8s within DODAS

INFN
Istituto Nazionale di Fisica Nucleare

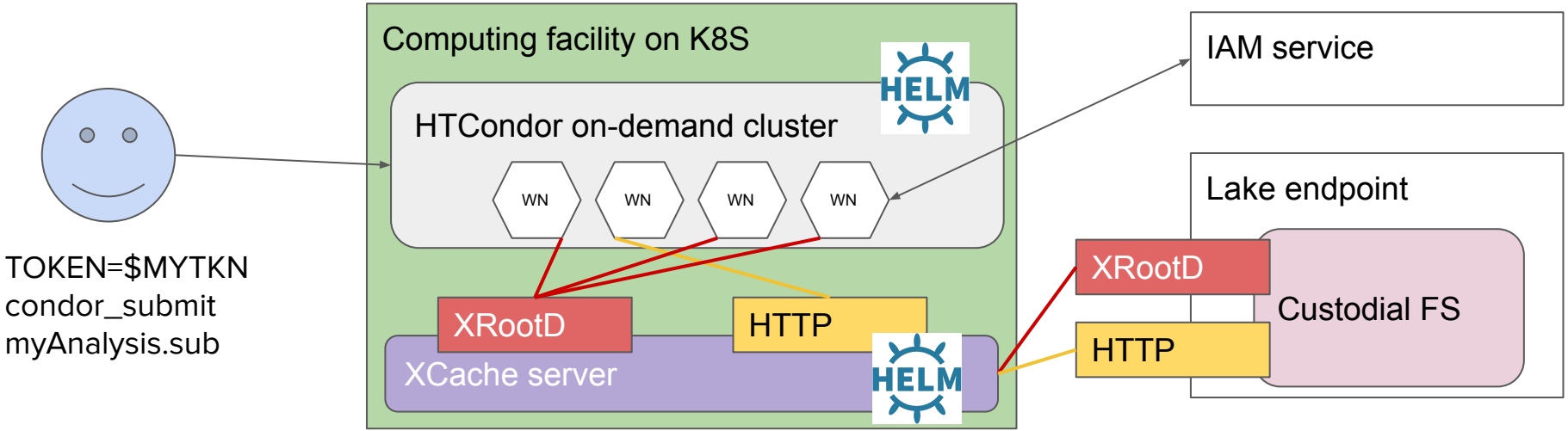
We use K8s+HELM as baseline solution for several DODAS use cases

- CMS HTCondor WNs integrated with central pool
- Jupyter+Spark on demand
- JupyterHUB and IAM integration
- **Analysis facilities on the landscape of data-lake testbeds**

Today we will focus mainly on the latter activity from the k8s point of view.

For more details regarding the DODAS approach see Daniele's presentation at pre-GDB K8s

Data-lake integration



Data-lake compute and data (cache) are provided via K8s+HELM and they include JWT (wlcg profile) based authN/Z (more details about the flow at [this DOMA pres]()

On the next slides we are going to focus on the k8s specific aspects for both (data and compute)

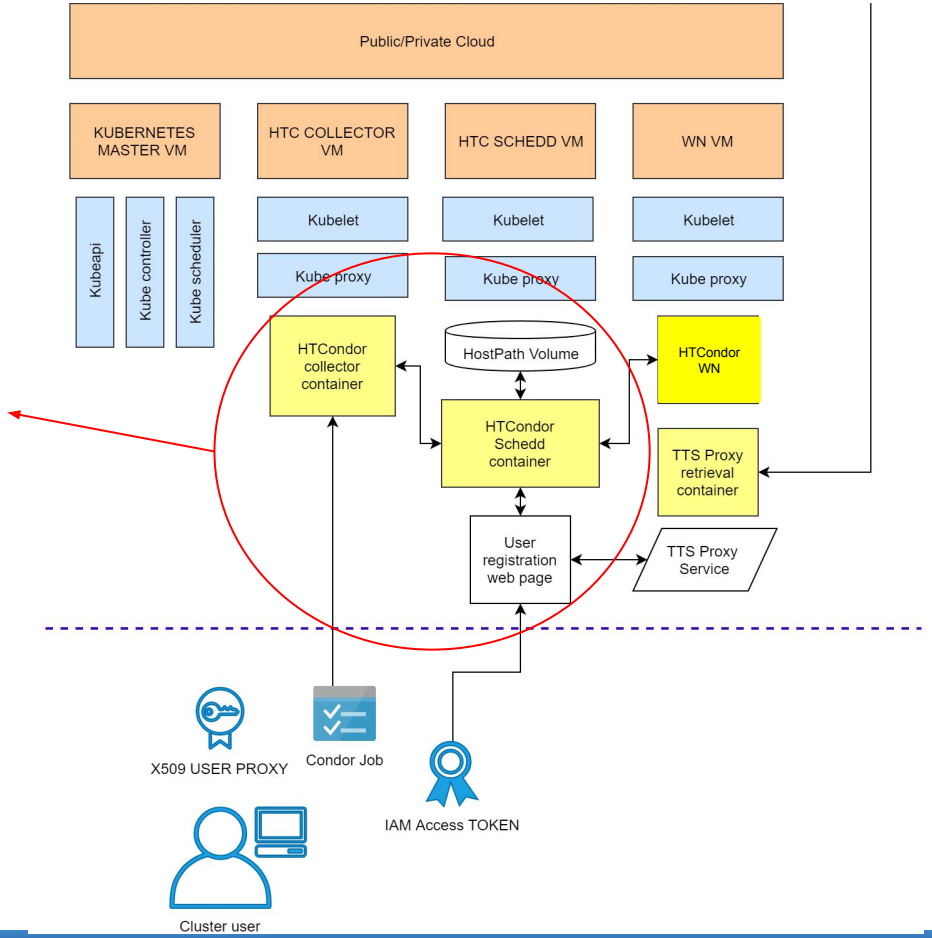
Compute

- Analysis Facility: we use DODAS to generate a **HTCondor pool managed by Kubernetes**
  - configuration fully managed through Helm charts
    - not specific to CMS
    - cvmfs and squid completely configurable at chart level
- Highly customizable and dynamic thanks to the use of ConfigMap for the configuration of every HTCondor component
  - Including batches federation (via flocking) and multi-tenancy setup
- **IAM** token used to retrieve a valid X509 proxy for daemon to daemon communication
  - This is to support “legacy” scenarios
- **Currently testing the HTCondor JWT management for data access tokens (credmon) and for fully integrated JWT authN/Z flow**

K8s HTCondor in a nutshell 1/2

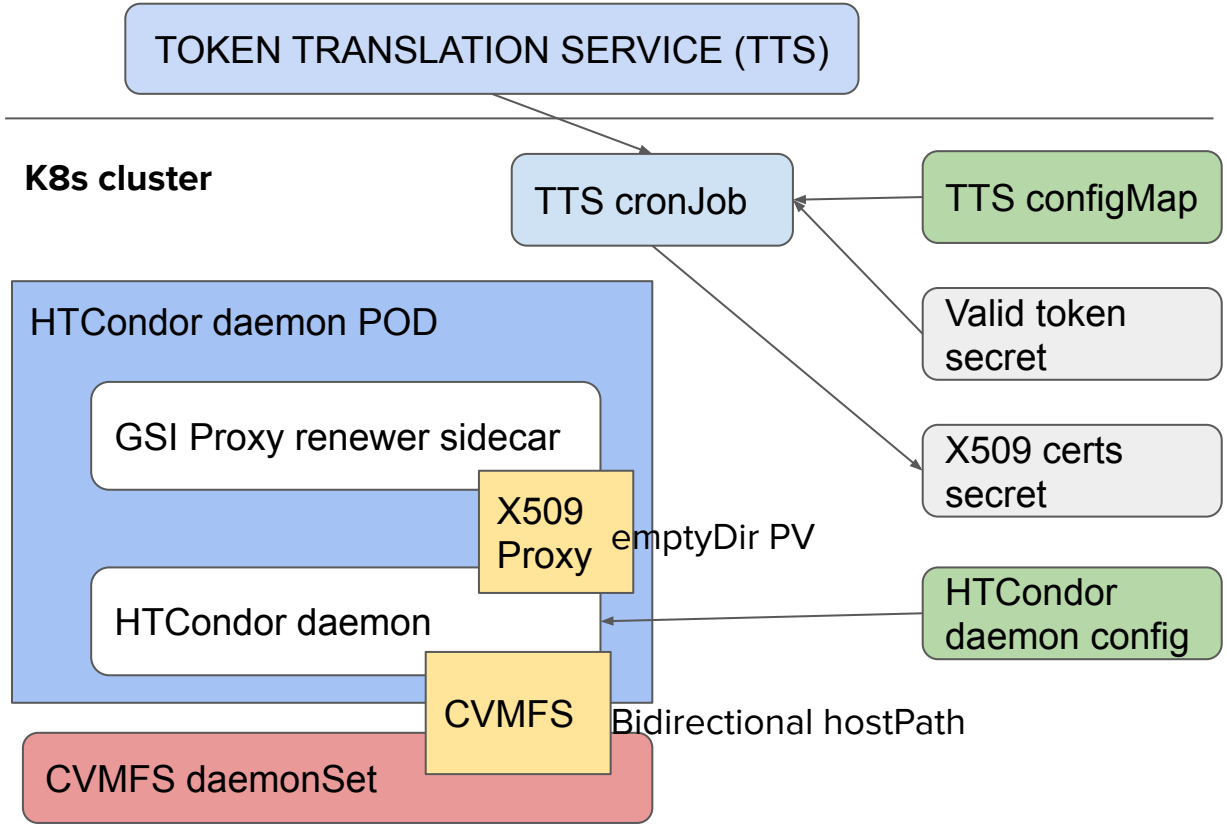
The deployment can be summarized as in the schema here

- The collector and schedd daemon run on 2 dedicated VMs with net:host for their particular needs
  - Using k8s node labels and affinities
- Sched spool is mounted via HostPath, but configurable through persistent volumes
- fully customizable through configMap and secrets



K8s HTCondor in a nutshell 2/2

Needed only until a direct authN/Z with token will be supported



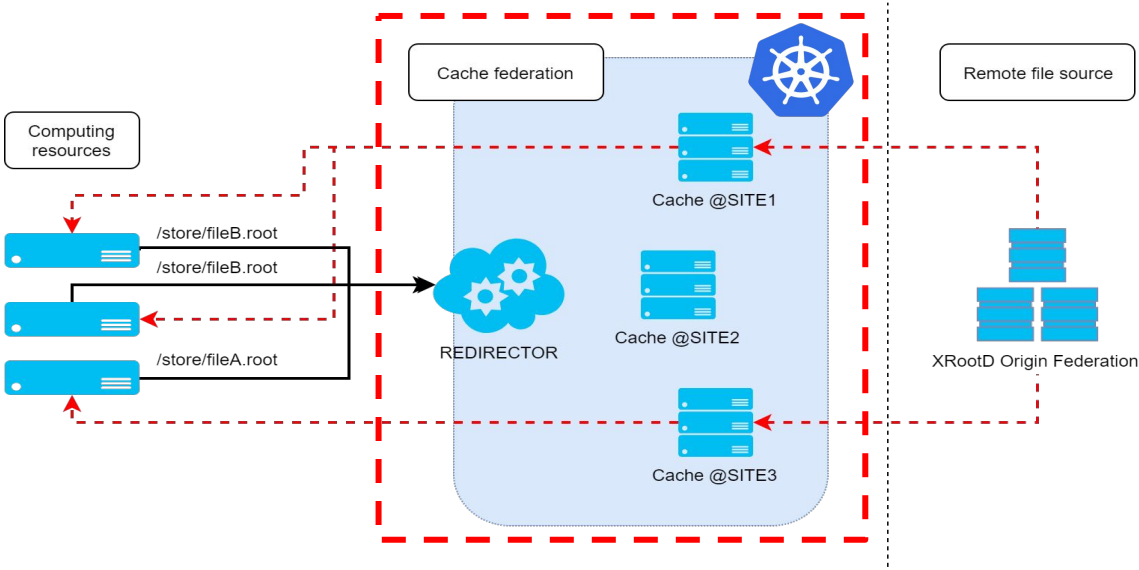
- Token to X509 proxy is automated
- Components configuration possible through ConfigMap (e.g. flocking)
- CVMFS can be mounted through hostPath

Storage

The data cache part leverages XRootD XCache technology to cache data near the WNs

- Configurable through Pod environment variables and/or configMap
- New cache servers can be scaled and automatically registered on the redirector
- Limit to one cache per VM and using net:host for performance reason

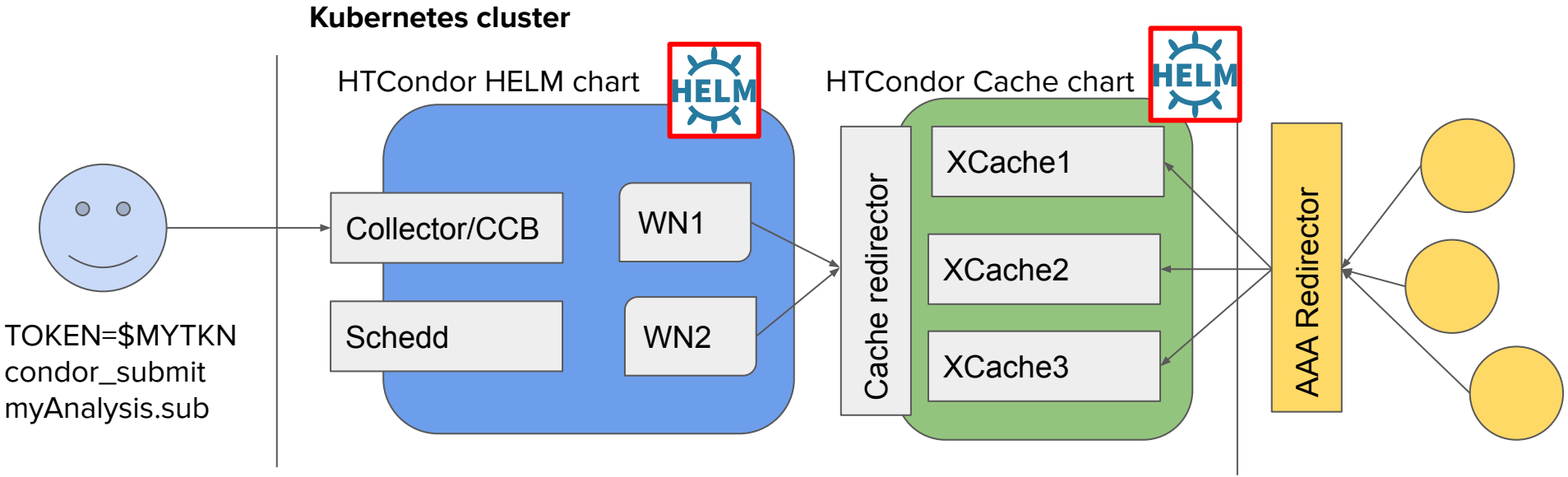
- Docker compose
- Caching on demand HELM chart
  - Currently moving from pods to daemonset over labelled nodes



In this case the cache is dynamically created but it could be just a pre-configured cache (e.g. at a lake, such as CNAF) to be shared by different standalone cluster

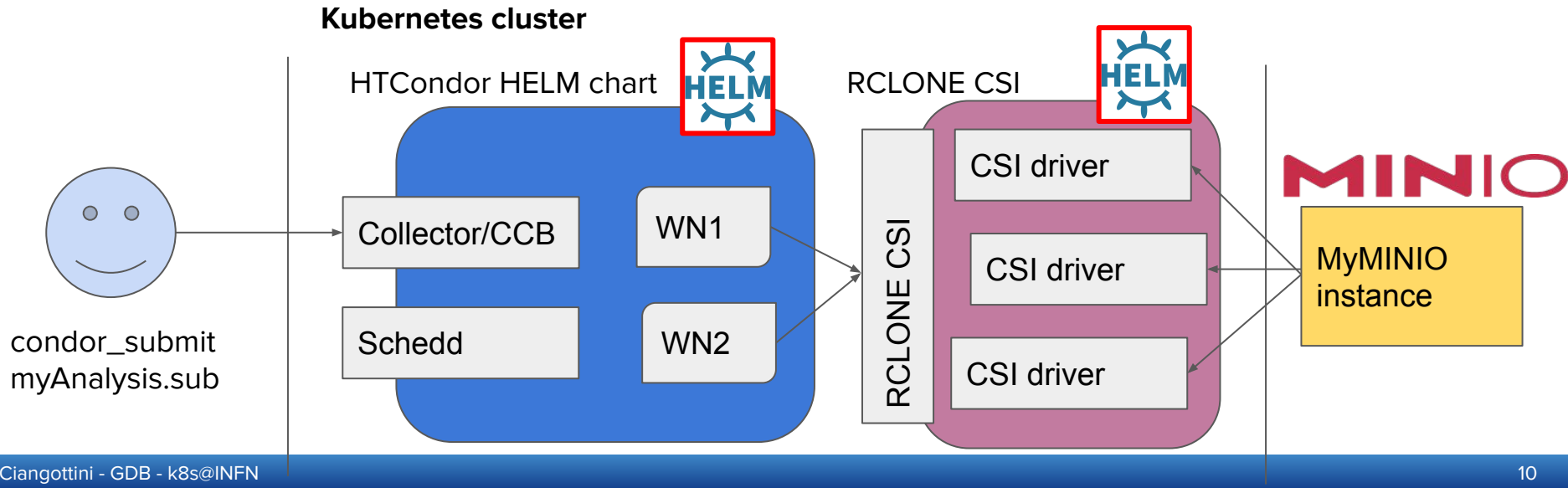
K8s based facility @CMS: the early tests

- We used this setup for a **demonstration over NanoAOD data in CMS**
- Planning for **a stable integration with real analysis** @ INFN-CNAF “ProtoLake”



Beyond CMS…

- A similar approach has been extended also to a FERMI-LAT experiment use case
  - No XRootD involved
  - **Posix through RClone (with caching on VFS) has been tested exposing buckets of our Object Storage (MINIO)**
- From K8s perspective we used a CSI produced to provide PersistentVolumes based on RClone
  - Very flexible approach to support a lot of backends that are compatible with it



Deploying IAM on K8S

IAM @ CNAF mainly deployed on K8S since 2018

~ 15 instances serving various communities

Streamlined deployment and operations

Tight integration with CI, rolling updates, self-healing

Kustomize for configuration management

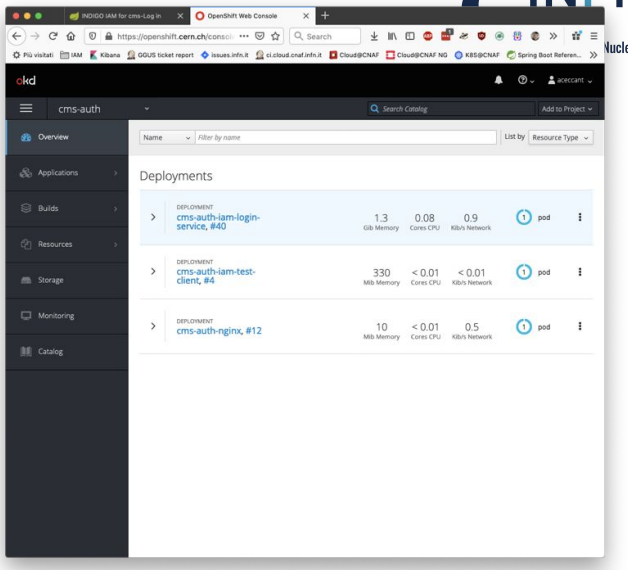
Makes bootstrapping a new IAM instance a matter of minutes

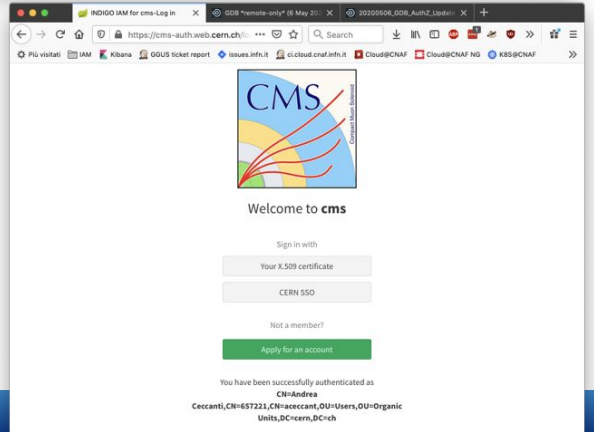
IAM @ CERN deployed on Openshift

Same approach with minor modifications on ingress controller management

Main source of headaches

OOM & scheduling problems when mis-managing deployment requests, limits and node resources capacity





StoRM & K8S

Istituto Nazionale di Fisica Nucleare

StoRM already containerised for development & testing (on plain POSIX FS)

Interest in using K8S to support service deployment/operations @ CNAF

Main problems under investigation: GPFS storage access

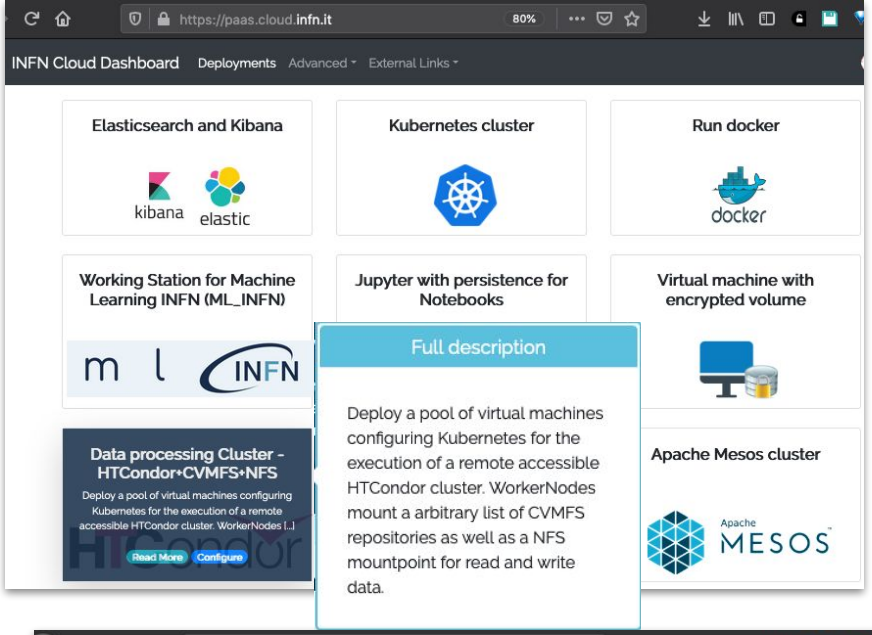
StoRM relies on GPFS APIs (for quota and tape interaction) not working in a container; plan is to refactor StoRM to only rely on POSIX APIs

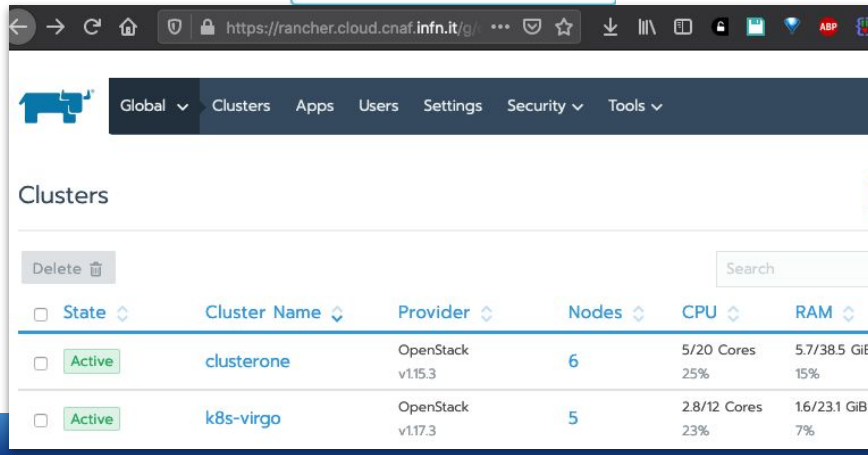
GPFS access from K8S

- As hostpath volumes
- Using the IBM Spectrum Scale Storage Enabler for Containers (SEC): i.e., connector that allows to use GPFS as a volume provider in K8S

K8s@INFN-Cloud

- **INFN Cloud** & **orchestration** of K8S clusters for different use cases
  - Deploy a **single K8S** 1.17.0 cluster
  - **Data processing cluster**: HTCondor+CVMFS+NFS, Spark+Jupyter
- In the framework of a **CERN-CNAF collaboration** regarding topics related to large scale infrastructures management
  - **K8S clusters on top of Openstack**
    - **Centrally managed** - KaaS through Rancher
    - Exploiting **GPUs/vGPUs**, benchmarking & performance testing (WIP)





Summary and future activities

Istituto Nazionale di Fisica Nucleare

Presented the usage of k8s to manage compute and data facilities in the data lake landscape

- The presented model is **compliant and integrated with INFN-Cloud**
- We also show use cases for **long running services @CNAF**
- Several active development topics that include:
  - **XCache k8s operator**
    - Manage XCache lifecycle with an operator
  - Managing accelerators
    - **GPUs**
    - **custom FPGA device plugins**
  - JupyterHub-IAM integration
    - interactive analyses at facilities
  - **Fine grained K8s authN/Z**
    - with OpenPolicyAgent and IAM

Backup

Finally… embedding in DODAS

Once we come to the DODAS based approach all this means moving toward Infrastructure as a code..

- To create and provision infrastructure deployments, automatically and repeatedly

Technically: the Helm is embedded in TOSCA template in our case

See DODAS doc for further details