Deployment dell'ambiente di calcolo per ML-INFN su INFN Cloud sfruttando le GPU

(Eventualmente partizionate)

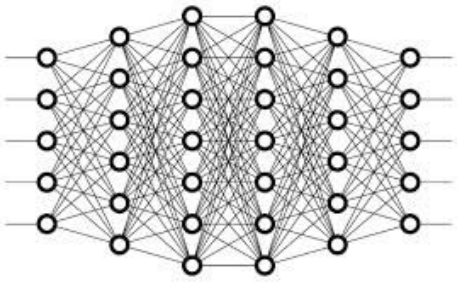
Speaker:
Gioacchino Vino

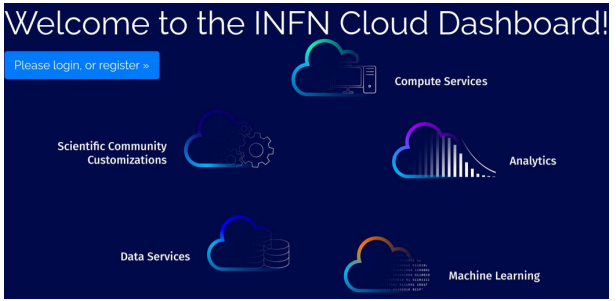
Authors: Marica Antonacci
Diego Ciangottini
Federico Fornari
Mauro Gattari
Daniele Spiga
Enrico Vianello
Gioacchino Vino

Workshop del Calcolo nell'INFN - Loano 23May23

Motivations

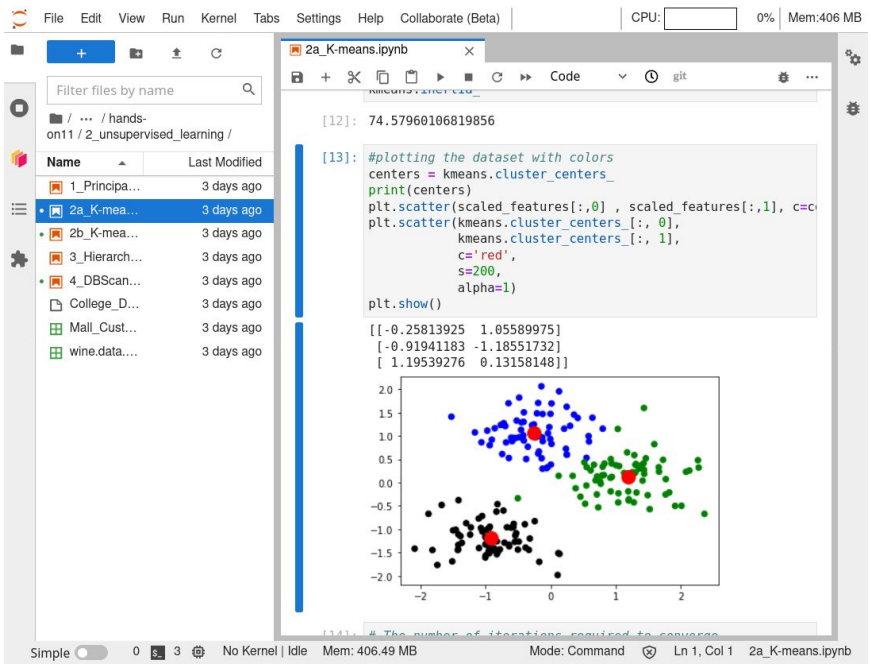
- Machine and Deep Learning algorithms
- High-performance parallel computing devices
- Remote resources usage
- Centralization of services





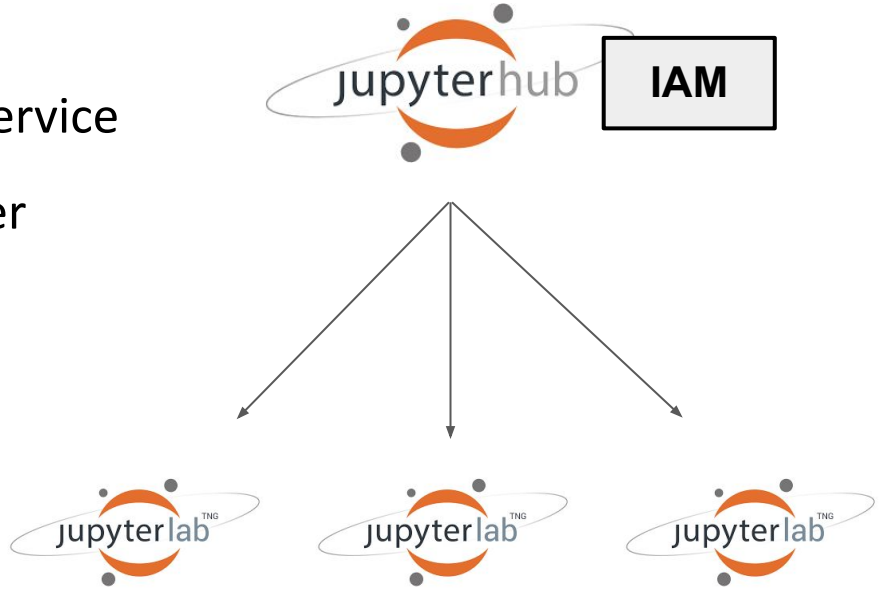
JupyterLab

- Web interface for writing and executing code
- Massively used in data-science and Machine Learning developments
- Extensions improve user experience
- Allows collaborative mode



JupyterHub

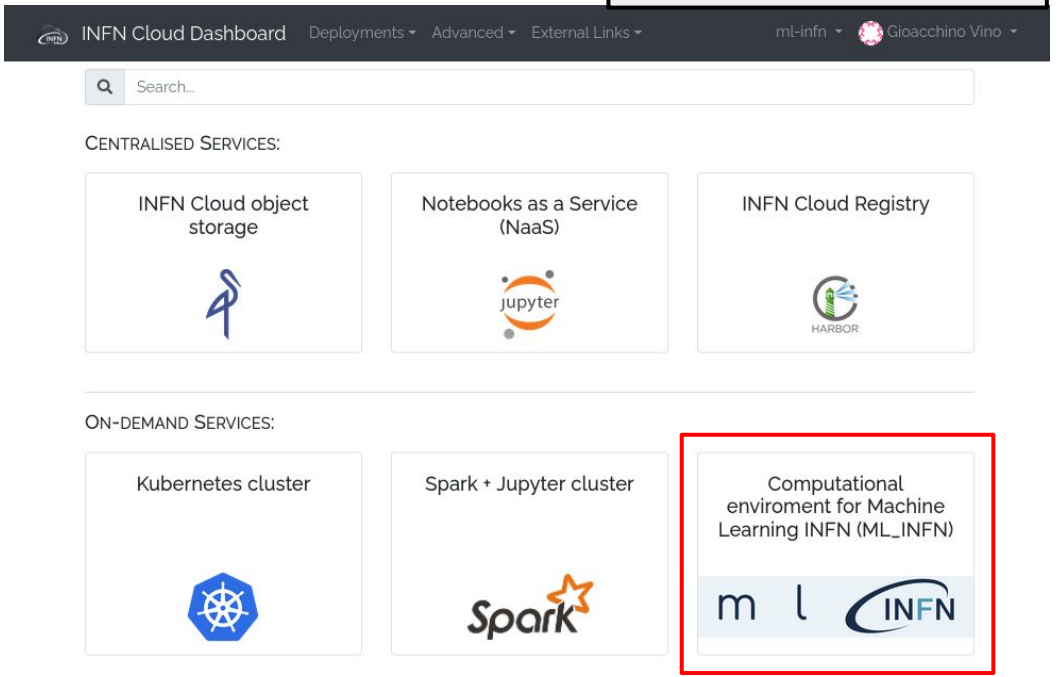
- Handles user authentication
- Extends JupyterLab to be a multi-user service
- Manages all accesses to different Jupyter instances
- Manages users and Jupyter instances



Talk Marica Antonacci

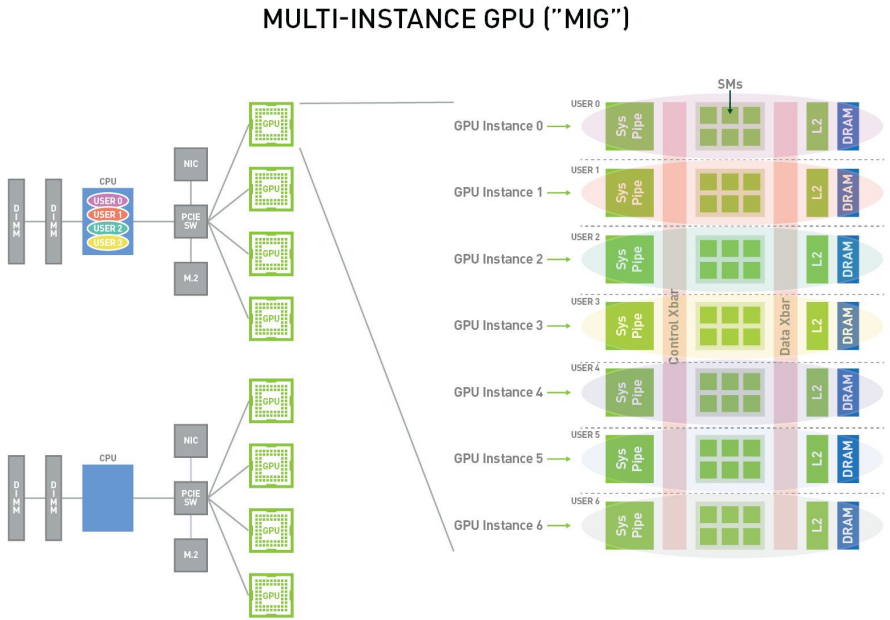
INFN-DataCloud Dashboard

- Accessible to my.cloud.infn.it
- Central Portal
- Centralised Services and On-Demand Services
- Portfolio
  - Object Storage
  - Notebook as a Service
  - Registry
  - Kubernetes cluster
  - Spark + Jupyter
  - Computational environment for Machine Learning INFN



NVIDIA GPU Partitioning

- Multi-Instance GPU (MIG) Technology
- NVIDIA architecture A30, A100 and H100
- up to 7 separate GPU Instances
- Ideal for workloads that do not fully saturate the GPU compute capacity



Computational environment for Machine Learning INFN

Goal:

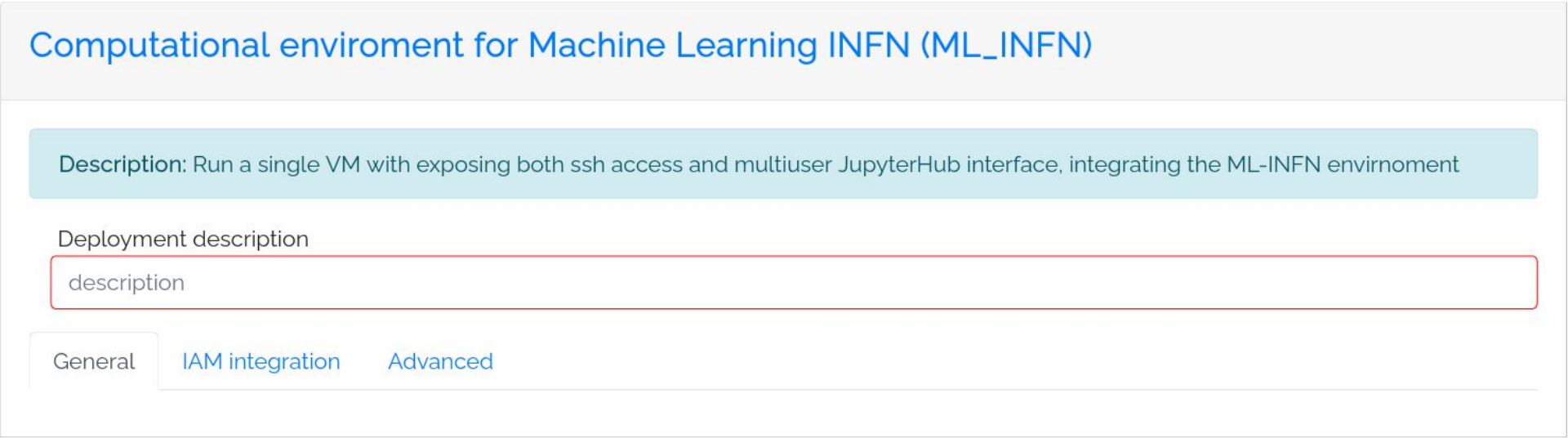
- Support users to configure the selected service

Phases:

- Collection of required information through the **INFN-Cloud Dashboard**
- Evaluation if required **resource are available** (VM)
- **Automatized configuration** of:
  - JupyterHub
  - GPU driver installation and partitioning

Computational environment for Machine Learning INFN

- Collection of information through the **INFN-Cloud Dashboard**:

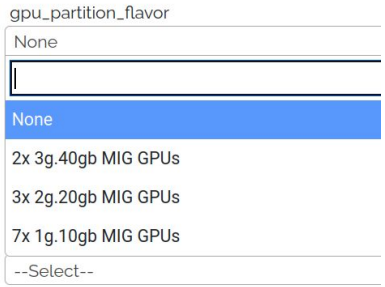


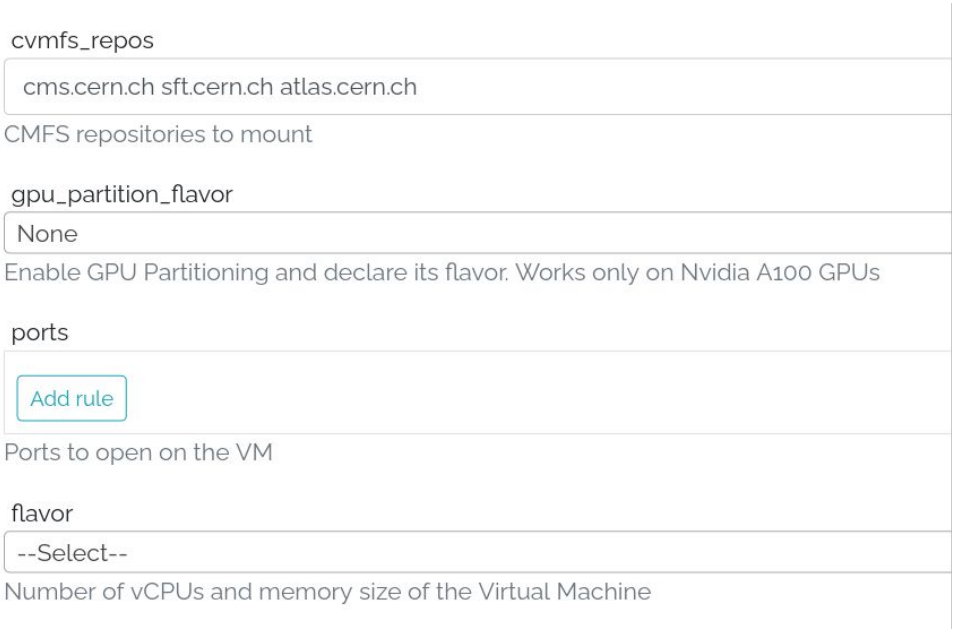
Computational environment for Machine Learning INFN

- Collection of information through the **INFN-Cloud Dashboard**:
  - Resource Monitoring
  - JupyterHub and JupyterLab

Computational environment for Machine Learning INFN

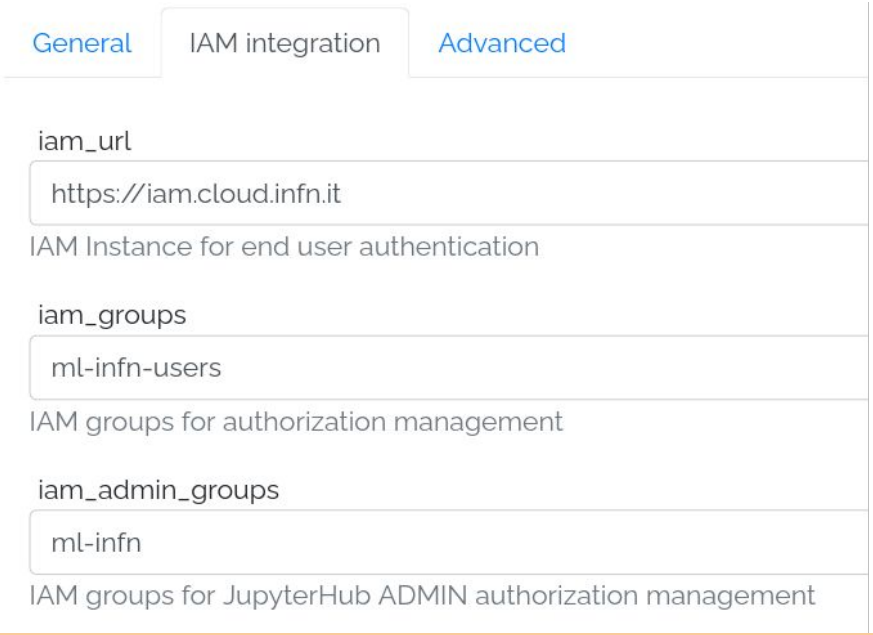
- Collection of information through the **INFN-Cloud Dashboard**:
  - Resource Monitoring
  - JupyterHub and JupyterLab
  - CVMFS Repository
  - VM flavor
  - GPU (potentially to partition)





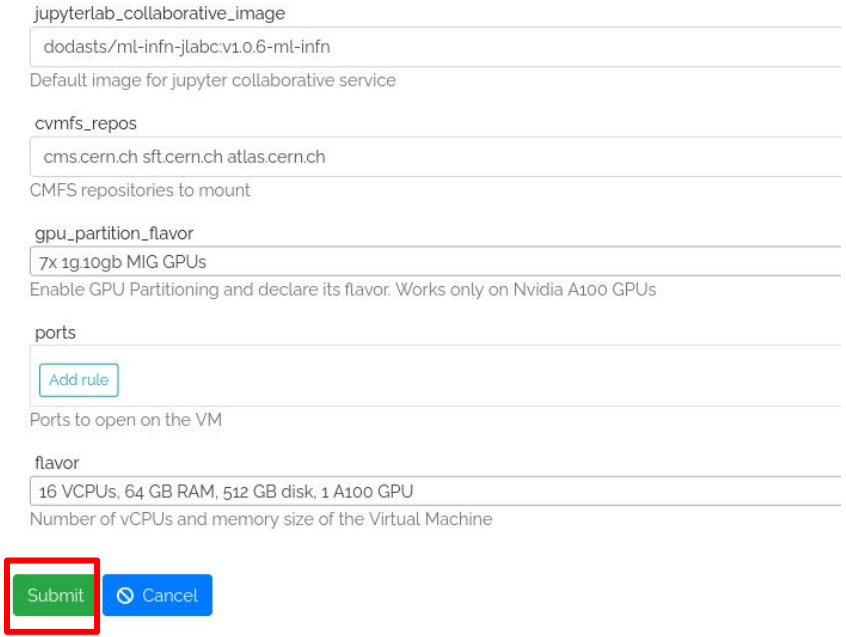
Computational environment for Machine Learning INFN

- Collection of information through the **INFN-Cloud Dashboard**:
  - Resource Monitoring
  - JupyterHub and JupyterLab
  - CVMFS Repository
  - VM flavor
  - GPU (potentially to partition)
  - IAM integration



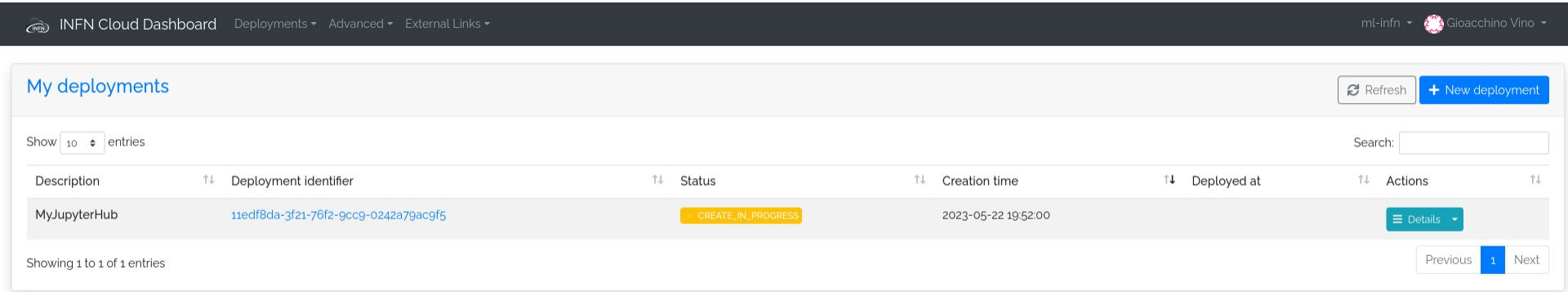
Computational environment for Machine Learning INFN

- Collection of information through the **INFN-Cloud Dashboard**:



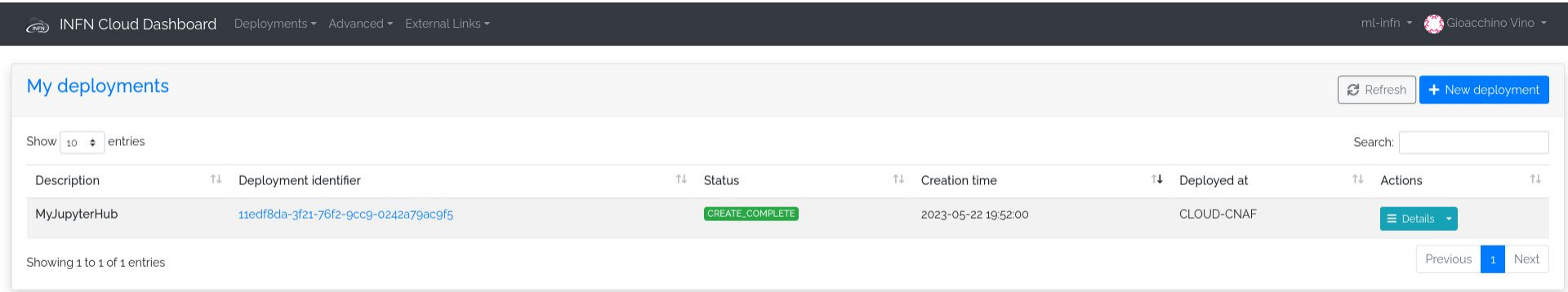
Computational environment for Machine Learning INFN

- Collection of required information through the **INFN-Cloud Dashboard**
- Evaluation if required **resource are available** (VM)
- **Automatized configuration** on the service of the Virtual Machine



Computational environment for Machine Learning INFN

- Collection of required information through the **INFN-Cloud Dashboard**
- Evaluation if required **resource are available** (VM)
- **Automatized configuration** on the service of the Virtual Machine



Computational environment for Machine Learning INFN

Deployment Description

11edf8da-3f21-76f2-9cc9-0242a79ac9f5

Description: MyJupyterHub

Overview | Input values | Output values

STATUS: CREATE_COMPLETE

CREATED AT: 2023-05-22 19:52:00

UPDATED AT: 2023-05-22 19:59:00

DEPLOYED AT: CLOUD-CNAF

11edf8da-3f21-76f2-9cc9-0242a79ac9f5 ← Back

Description: MyJupyterHub

Overview | Input values | Output values

additional_description: MyJupyterHub

cvmfs_repos: cms.cern.ch sft.cern.ch atlas.cern.ch

disk_size: 512 GB

enable_monitoring: false

gpu_model:

gpu_partition_flavor: None

iam_admin_groups: ml-infn

iam_groups: ml-infn-users

iam_subject: 0e574d04-e61a-48ac-82aa-7e4eeaeeda88

iam_url: https://iam.cloud.infn.it

jupyter_images: dodasts/ml-infn-lab:v1.0.6-ml-infn

jupyter_use_gpu: false

jupyterlab_collaborative: false

jupyterlab_collaborative_image: dodasts/ml-infn-jlabc:v1.0.6-ml-infn

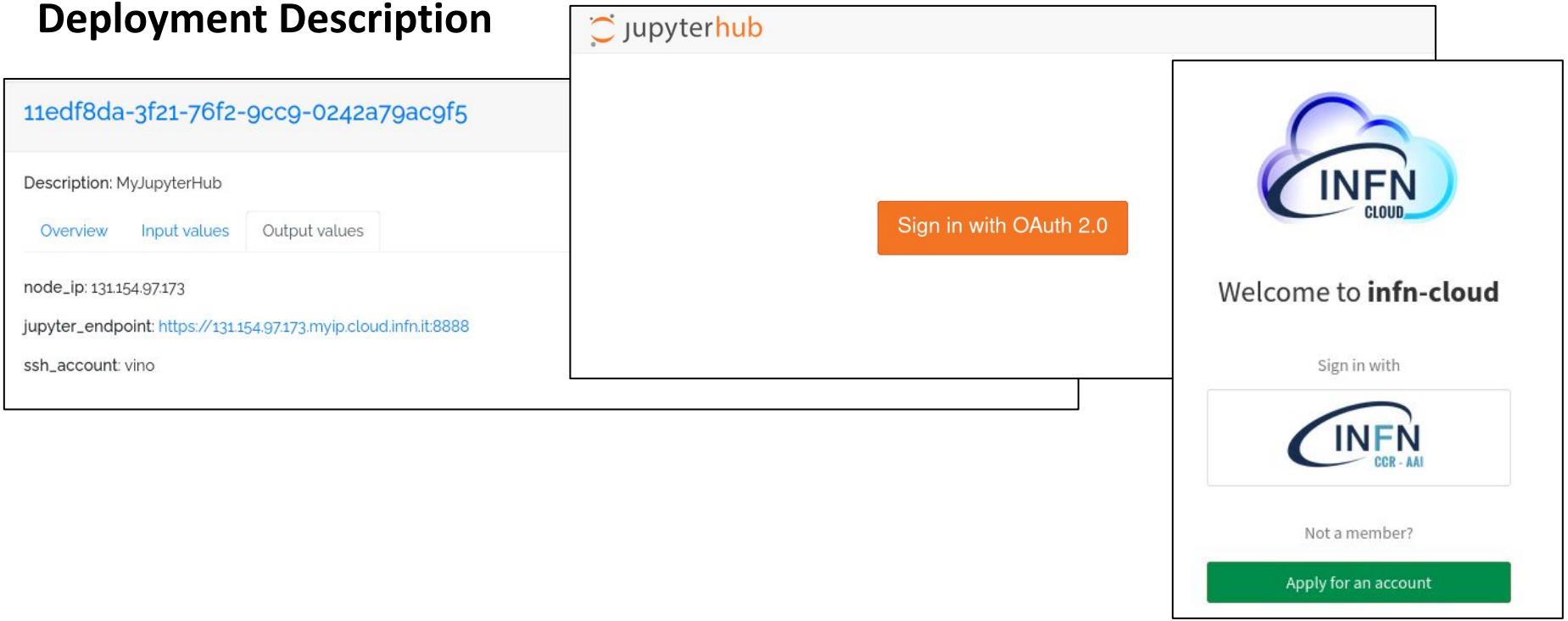
jupyterlab_collaborative_use_gpu: false

mem_size: 64 GB

num_cpus: 8

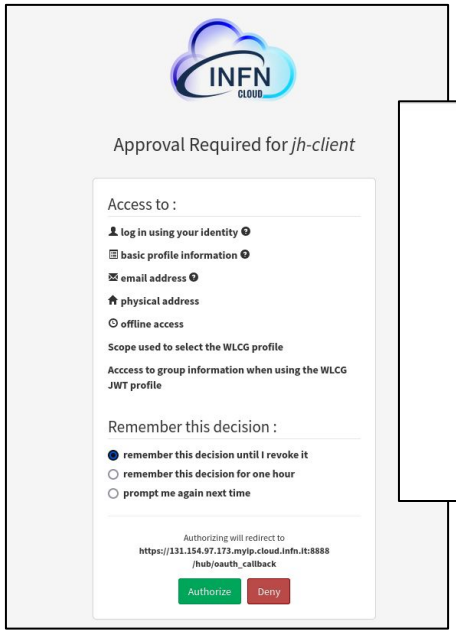
Computational environment for Machine Learning INFN

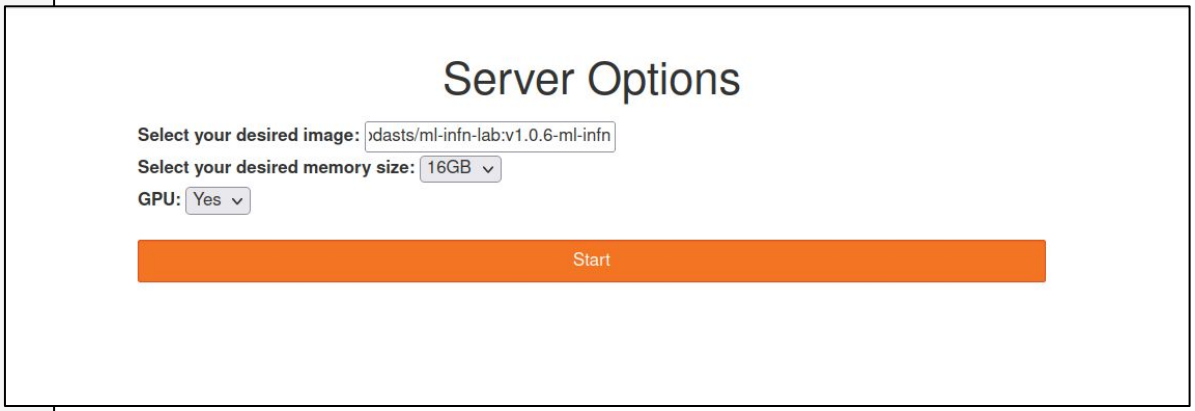
Deployment Description



Computational environment for Machine Learning INFN

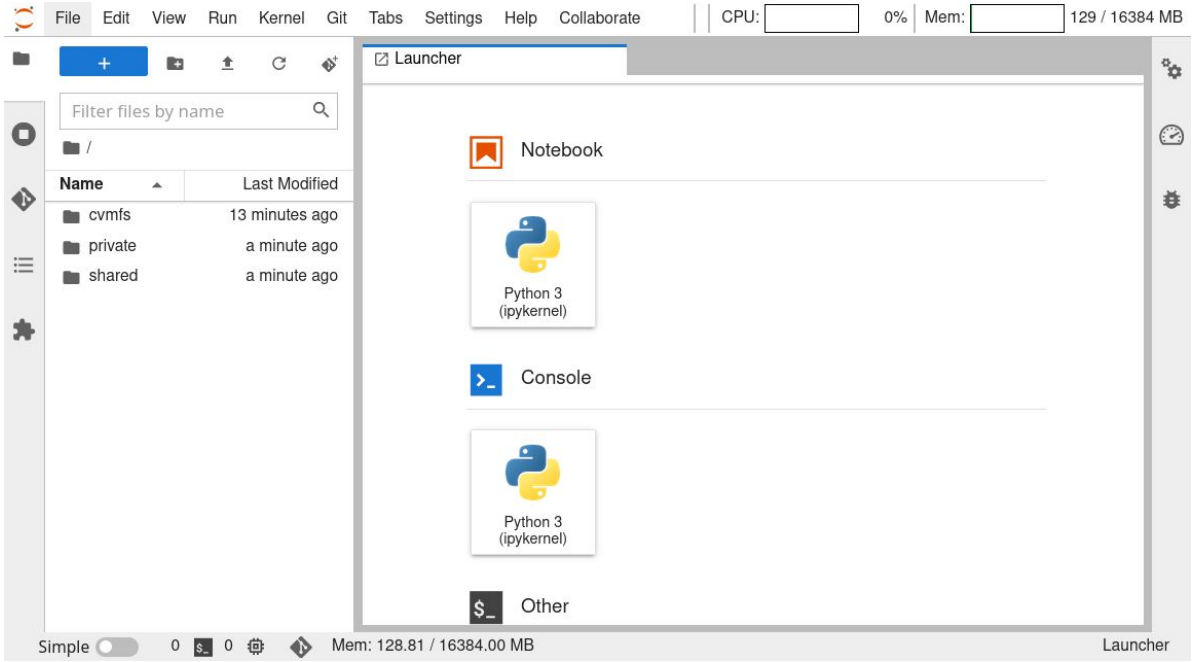
Deployment Description





Computational environment for Machine Learning INFN

Deployment Description



Under the hood

- Used Technologies:
  - TOSCA template and types
  - Ansible roles
  - PaaS-Orchestrator
  - Multiple Cloud Providers
- A generic and unified Ansible role has been development to take care GPU driver installation and partitioning in different services

Conclusions

- The GPU Partitioning improves the past version of the service
- Starting of a single A100, 7 virtual GPUs can be generated
- More users can take advantage from hardware-accelerated GPU
- MIG technology makes GPU usage more efficient respect past, even though users own a GPU with less performance

THANKS FOR YOUR ATTENTION

&