

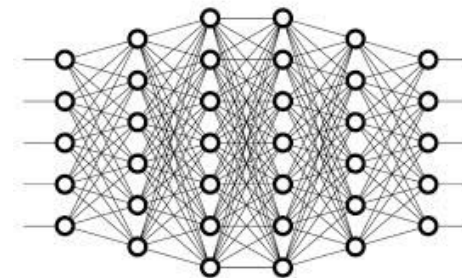
# Deployment dell'ambiente di calcolo per ML-INFN su INFN Cloud sfruttando le GPU (Eventualmente partizionate)

Speaker:  
Giacchino Vino

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

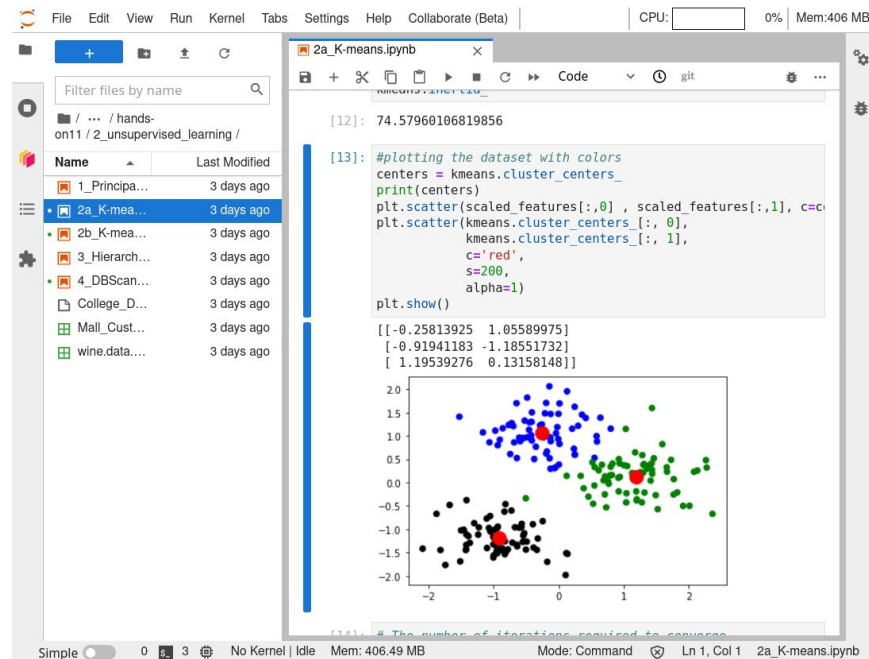
## Motivations

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



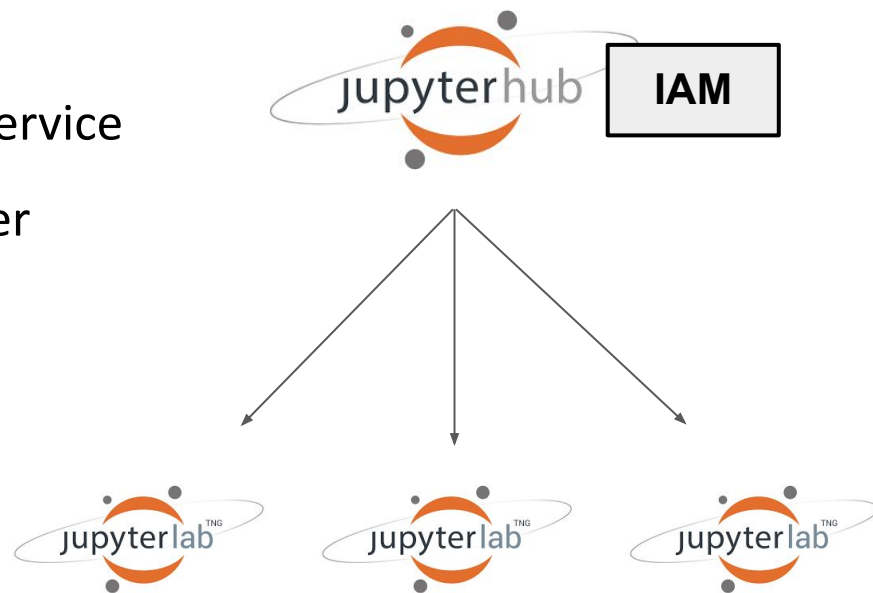
## JupyterLab

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



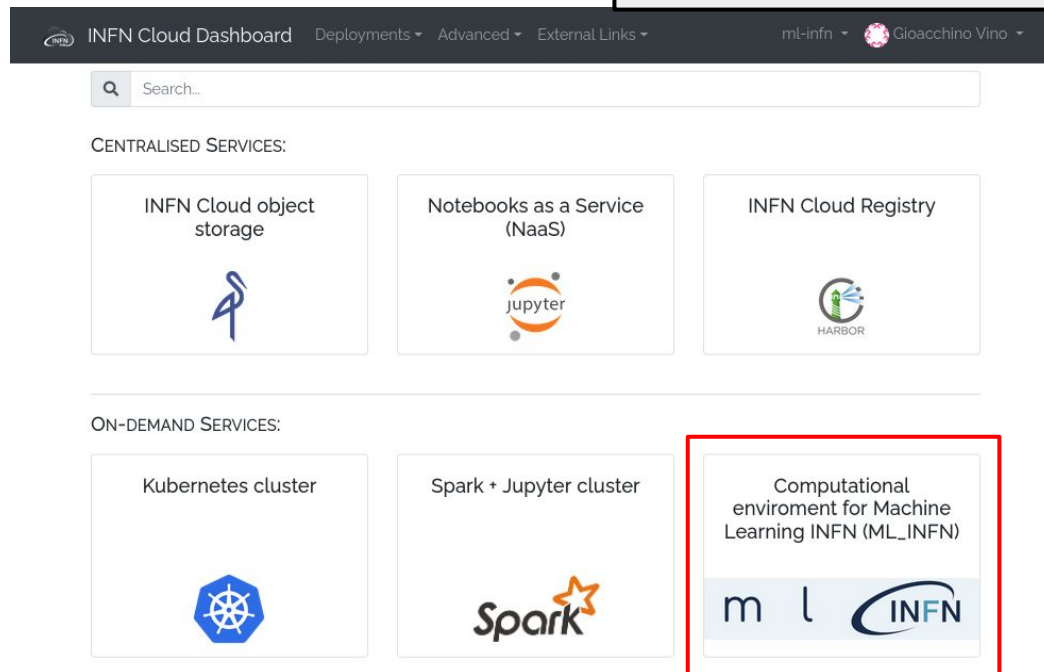
## JupyterHub

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



## INFN-DataCloud Dashboard

- Accessible to [my.cloud.infn.it](https://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



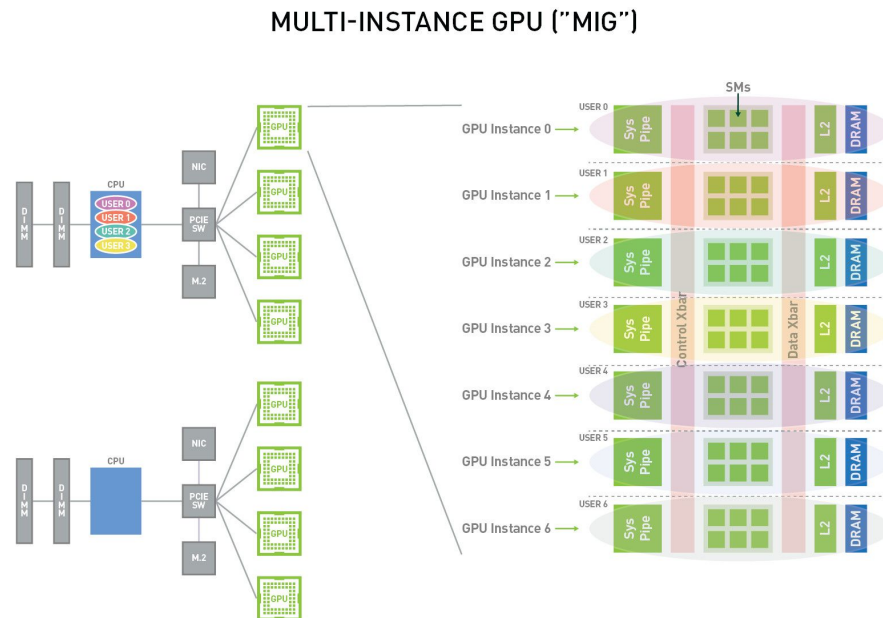
The screenshot shows the INFN Cloud Dashboard interface. At the top, there is a navigation bar with the INFN Cloud logo, the text "INFN Cloud Dashboard", and several menu items: "Deployments", "Advanced", and "External Links". On the right side of the navigation bar, there are user-related elements: "ml-infn" and a profile picture of "Giacchino Vino". Below the navigation bar is a search bar with the placeholder text "Search...". The main content area is divided into two sections: "CENTRALISED SERVICES:" and "ON-DEMAND SERVICES:". Under "CENTRALISED SERVICES:", there are three service cards: "INFN Cloud object storage" with a blue icon, "Notebooks as a Service (NaaS)" with the Jupyter logo, and "INFN Cloud Registry" with the Harbor logo. Under "ON-DEMAND SERVICES:", there are three service cards: "Kubernetes cluster" with the Kubernetes logo, "Spark + Jupyter cluster" with the Spark logo, and "Computational environment for Machine Learning INFN (ML\_INFN)" which is highlighted with a red border. This last card features the letters "m l" and the INFN logo.

## 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 (ML-INFN)

Description: Run a single VM with exposing both ssh access and multiuser JupyterHub interface, integrating the ML-INFN environment

Deployment description

General

IAM integration

Advanced



## Computational environment for Machine Learning INFN

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

enable\_monitoring

false

Enable/disable monitoring

jupyter\_images

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

Default image for jupyter server

jupyter\_use\_gpu

true

Enable GPU utilization on jupyter

jupyterlab\_collaborative

false

enable the jupyter collaborative service

jupyterlab\_collaborative\_use\_gpu

false

enable the GPU on jupyter collaborative service

jupyterlab\_collaborative\_image

dodasts/ml-infn-jlab:v1.0.6-ml-infn

Default image for jupyter collaborative service

## 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)

gpu\_partition\_flavor

None

None

2x 3g.40gb MIG GPUs

3x 2g.20gb MIG GPUs

7x 1g.10gb MIG GPUs

--Select--

cvnfs\_repos

cms.cern.ch sft.cern.ch atlas.cern.ch

CMFS repositories to mount

gpu\_partition\_flavor

None

Enable GPU Partitioning and declare its flavor. Works only on Nvidia A100 GPUs

ports

Add rule

Ports to open on the VM

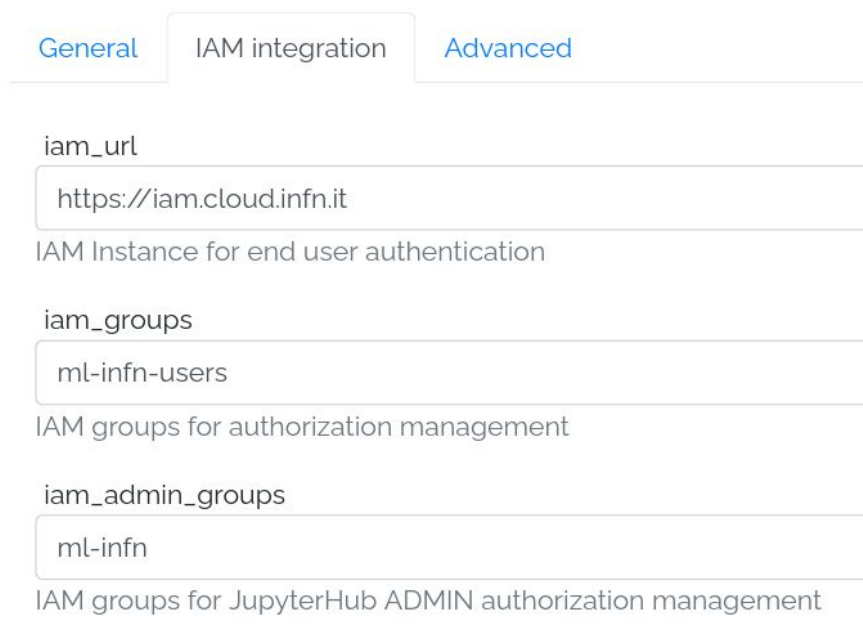
flavor

--Select--

Number of vCPUs and memory size of the Virtual Machine

## 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



The screenshot shows the 'IAM integration' tab of the dashboard. It contains three configuration sections:

- iam\_url**: A text input field containing the URL `https://iam.cloud.infn.it`. Below it is the label 'IAM Instance for end user authentication'.
- iam\_groups**: A text input field containing `ml-infn-users`. Below it is the label 'IAM groups for authorization management'.
- iam\_admin\_groups**: A text input field containing `ml-infn`. Below it is the label 'IAM groups for JupyterHub ADMIN authorization management'.

## Computational environment for Machine Learning INFN

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

jupyterlab\_collaborative\_image  
dodasts/ml-infn-jlab:v1.0.6-ml-infn  
Default image for jupyter collaborative service

cvmfs\_repos  
cms.cern.ch sft.cern.ch atlas.cern.ch  
CMFS repositories to mount

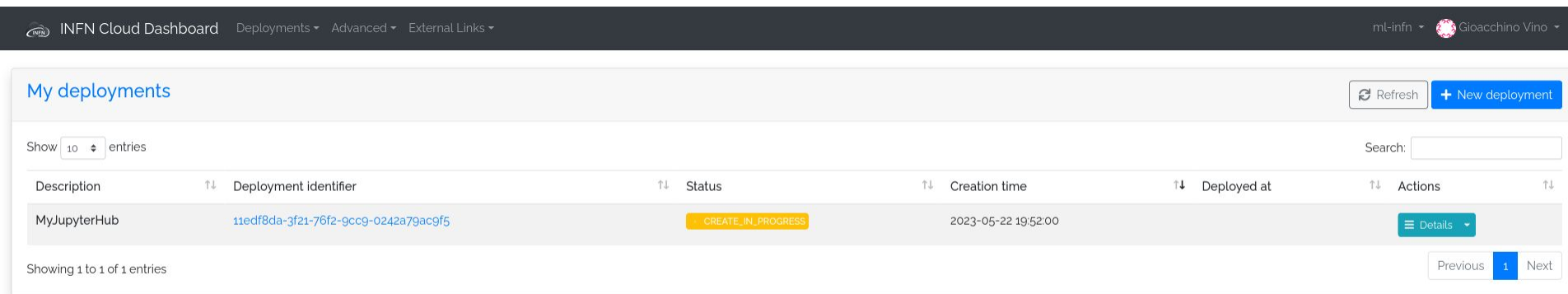
gpu\_partition\_flavor  
7x 1g.10gb MIG GPUs  
Enable GPU Partitioning and declare its flavor. Works only on Nvidia A100 GPUs

ports  
  
Ports to open on the VM

flavor  
16 VCPUs, 64 GB RAM, 512 GB disk, 1 A100 GPU  
Number of vCPUs and memory size 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



The screenshot shows the INFN Cloud Dashboard interface. At the top, there is a navigation bar with "INFN Cloud Dashboard" and several dropdown menus: "Deployments", "Advanced", and "External Links". On the right side of the navigation bar, there is a user profile for "Giacchino Vino".

The main content area is titled "My deployments". It features a "Refresh" button and a "+ New deployment" button. Below this, there is a "Show 10 entries" dropdown and a search input field.

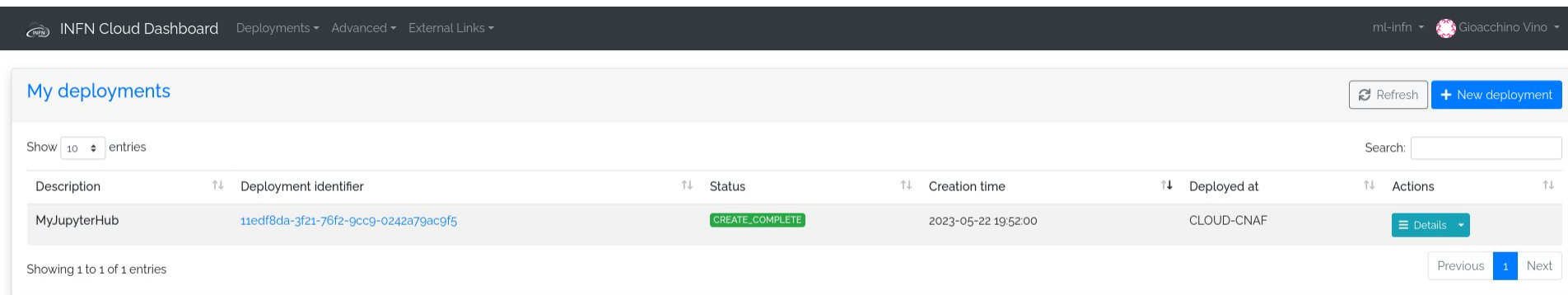
The central part of the dashboard is a table with the following columns: "Description", "Deployment identifier", "Status", "Creation time", "Deployed at", and "Actions".

Description	Deployment identifier	Status	Creation time	Deployed at	Actions
MyJupyterHub	11edf8da-3f21-76f2-9cc9-0242a79ac9f5	CREATE_IN_PROGRESS	2023-05-22 19:52:00		Details

At the bottom of the table, it says "Showing 1 to 1 of 1 entries". There are also "Previous", "1", and "Next" navigation buttons.

## 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



The screenshot shows the INFN Cloud Dashboard interface. At the top, there is a navigation bar with "INFN Cloud Dashboard" and several menu items: "Deployments", "Advanced", and "External Links". On the right side of the navigation bar, there is a user profile for "ml-infn" and "Giacchino Vino".

The main content area is titled "My deployments". It features a "Refresh" button and a "+ New deployment" button. Below this, there is a "Show 10 entries" dropdown and a search input field.

The central part of the dashboard is a table with the following columns: "Description", "Deployment identifier", "Status", "Creation time", "Deployed at", and "Actions".

Description	Deployment identifier	Status	Creation time	Deployed at	Actions
MyJupyterHub	11edf8da-3f21-76f2-9cc9-0242a79ac9f5	CREATE_COMPLETE	2023-05-22 19:52:00	CLOUD-CNAF	Details

At the bottom of the table, it says "Showing 1 to 1 of 1 entries". There are also "Previous", "1", and "Next" navigation buttons.

# 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-jlabcv1.0.6-ml-infn

jupyterlab\_collaborative\_use\_gpu: false

mem\_size: 64 GB

num\_cpus: 8

# Computational environment for Machine Learning INFN

## Deployment Description

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

Description: MyJupyterHub

Overview Input values Output values

node\_ip: 131.154.97.173

jupyter\_endpoint: <https://131.154.97.173.myip.cloud.infn.it:8888>

ssh\_account: vino

 jupyterhub

Sign in with OAuth 2.0



Welcome to **infn-cloud**

Sign in with



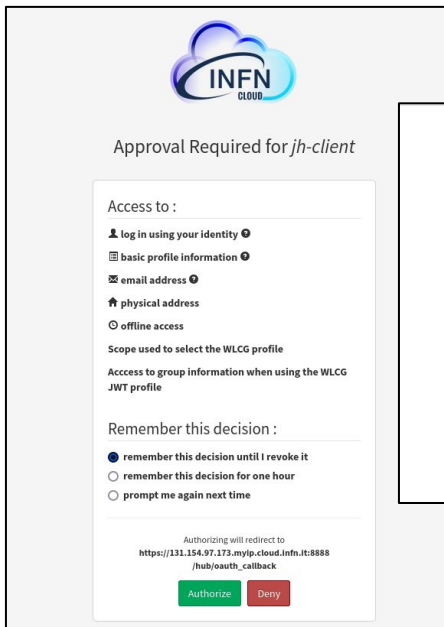
Not a member?

Apply for an account



# Computational environment for Machine Learning INFN

## Deployment Description



Approval Required for *jh-client*

Access to :

- log in using your identity
- basic profile information
- email address
- physical address
- offline access

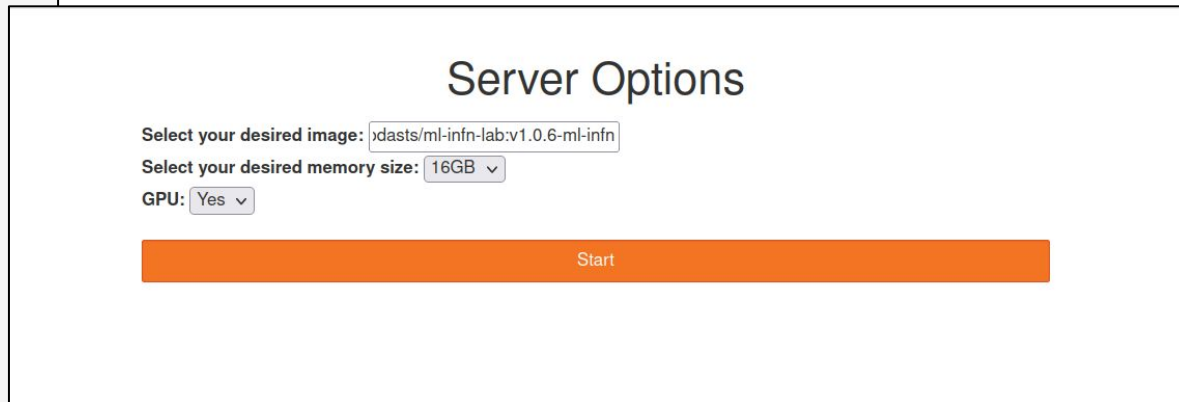
Scope used to select the WLCG profile

Access to group information when using the WLCG JWT profile

Remember this decision :

- remember this decision until I revoke it
- remember this decision for one hour
- prompt me again next time

Authorizing will redirect to  
[https://131.154.97.173.myip.cloud.infn.it:8888/hub/oauth\\_callback](https://131.154.97.173.myip.cloud.infn.it:8888/hub/oauth_callback)



### Server Options

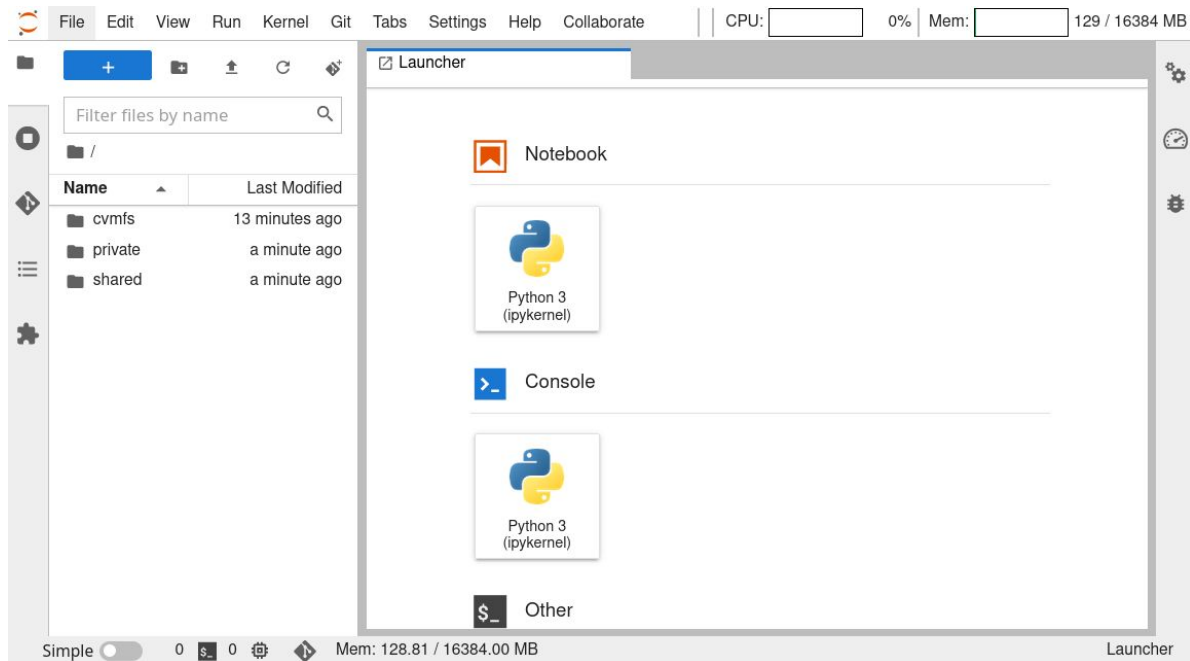
Select your desired image:

Select your desired memory size:

GPU:

# 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 developed to take care of 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**

**EXPLORE**

**&**

