Deployment di un ambiente di calcolo basato su JupyterHub con accesso a GPU tramite la PaaS di INFN-DataCloud

Gioacchino Vino

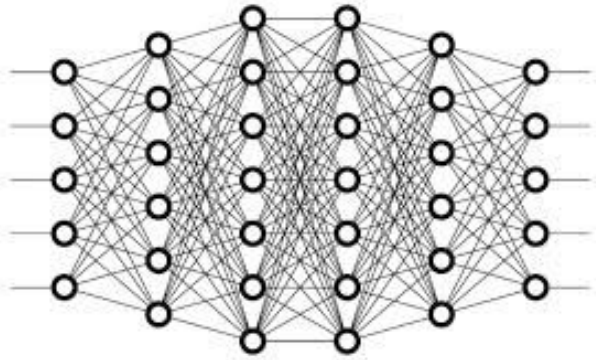
INFN

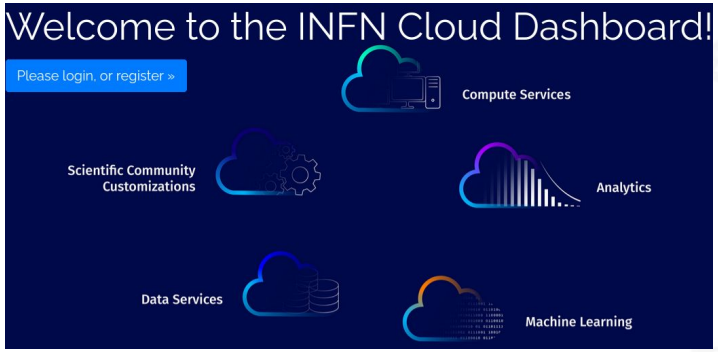
WORKSHOP GARR 2023

NET MAKERS

Motivations

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

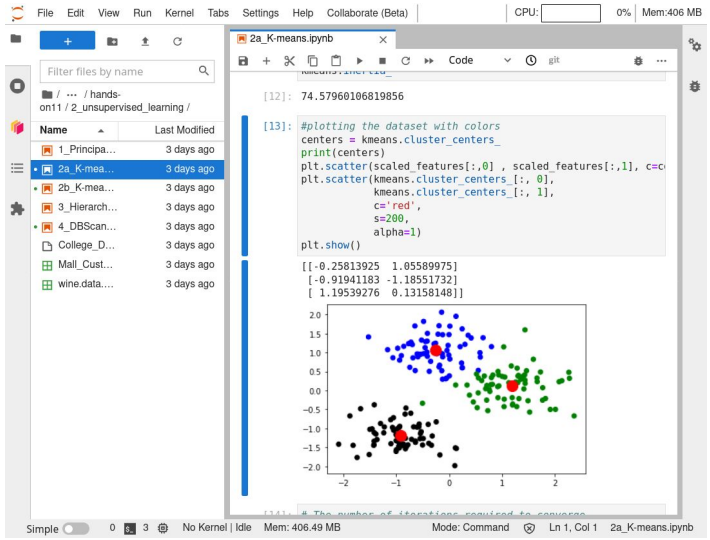




Jupyter

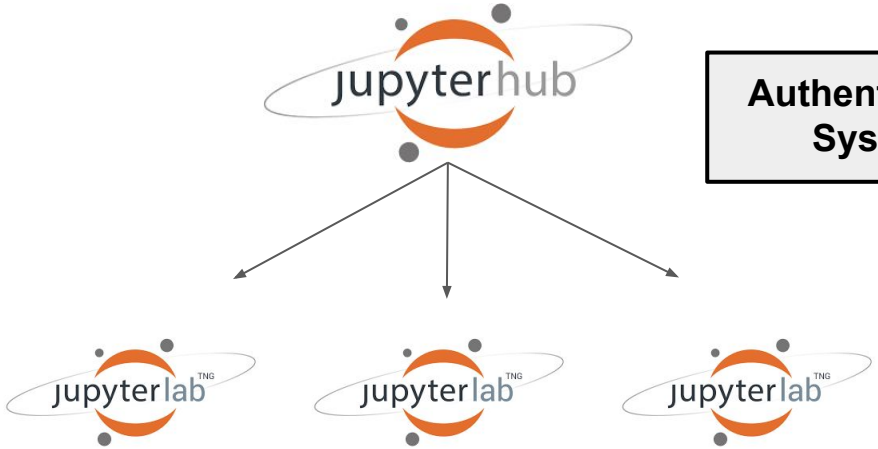
JupyterLab

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



JupyterHub

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

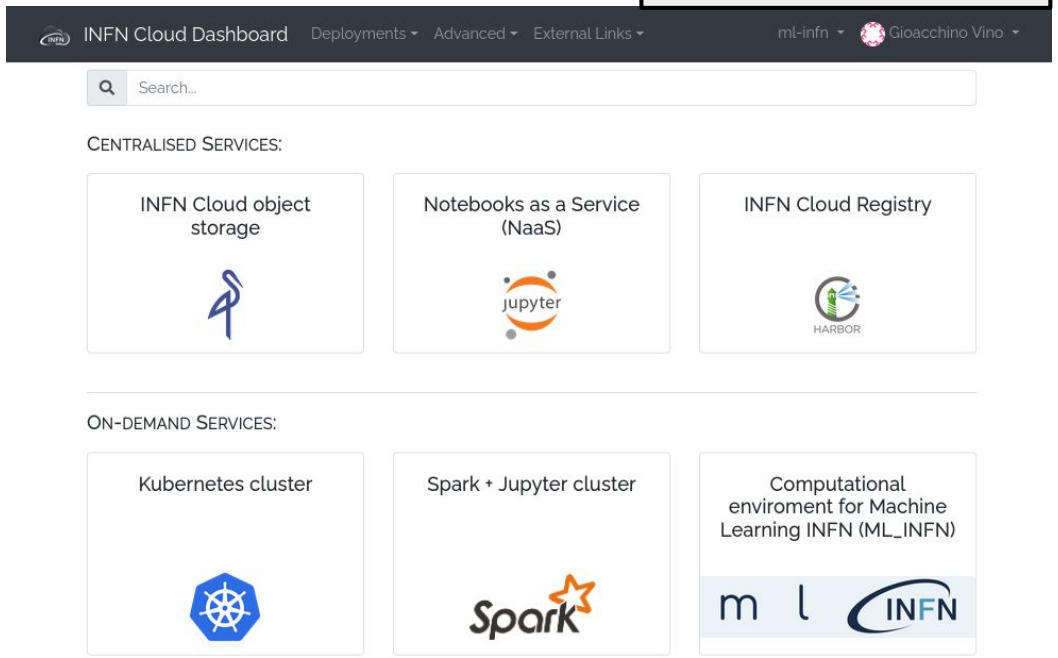


Authentication System

INFN-DataCloud Dashboard

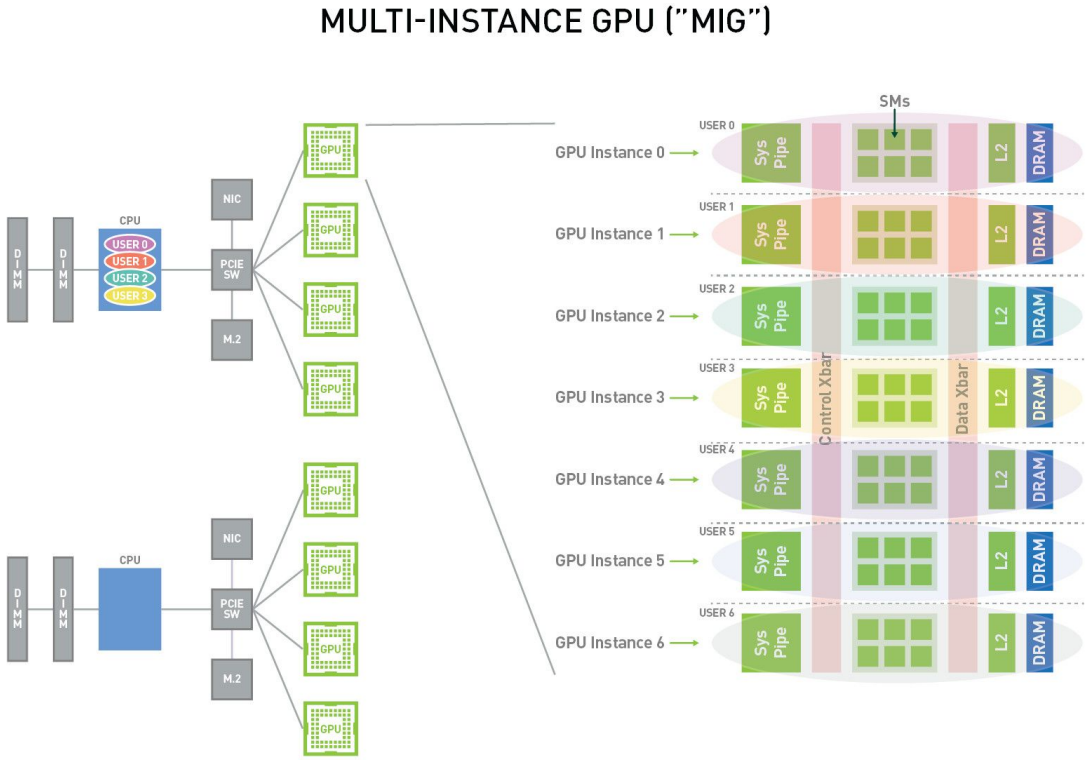
- Accessible to my.cloud.infn.it
- Central Portal
- Centralised Services and On-Demand Services
- Portfolio
  - Object Storage
  - Notebook as a Service
  - Registry
  - Kubernetes cluster
  - Spark + Jupyter
  - Computational environment for Machine Learning INFN

Talk Barbara Martelli



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



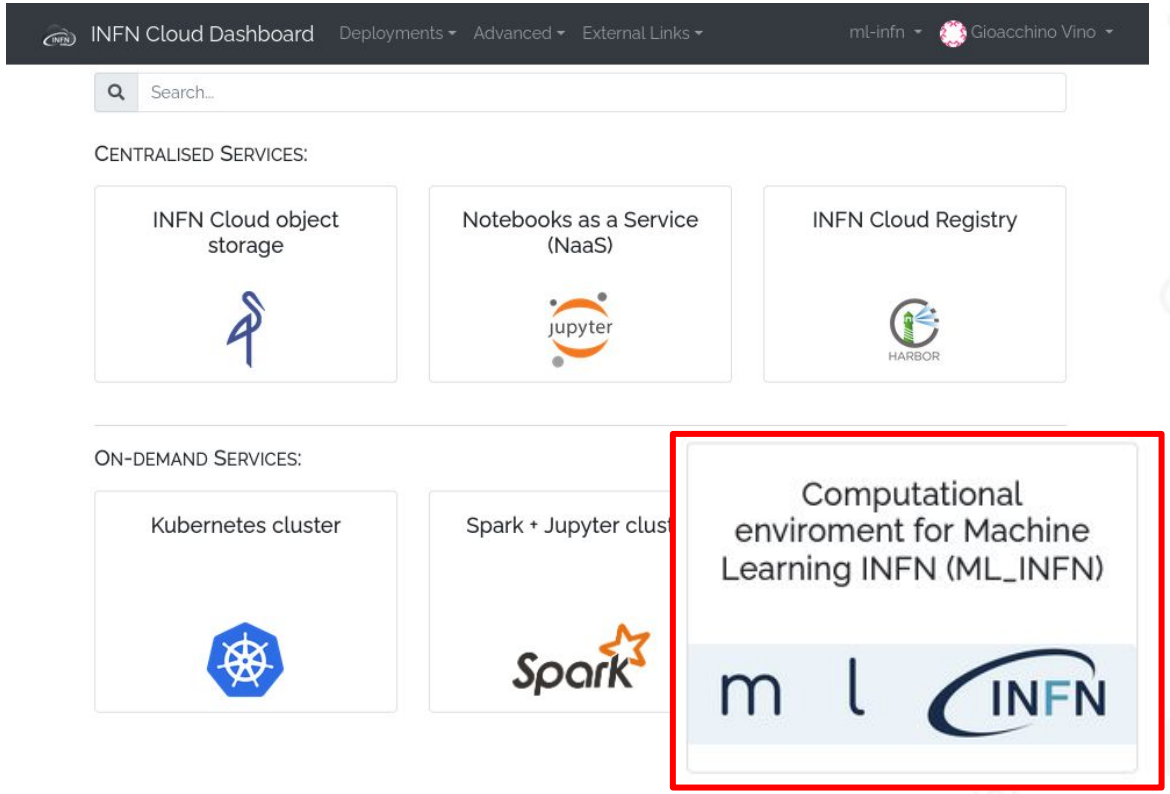
Computational environment for Machine Learning INFN

Goal:

- Support users to configure the selected service

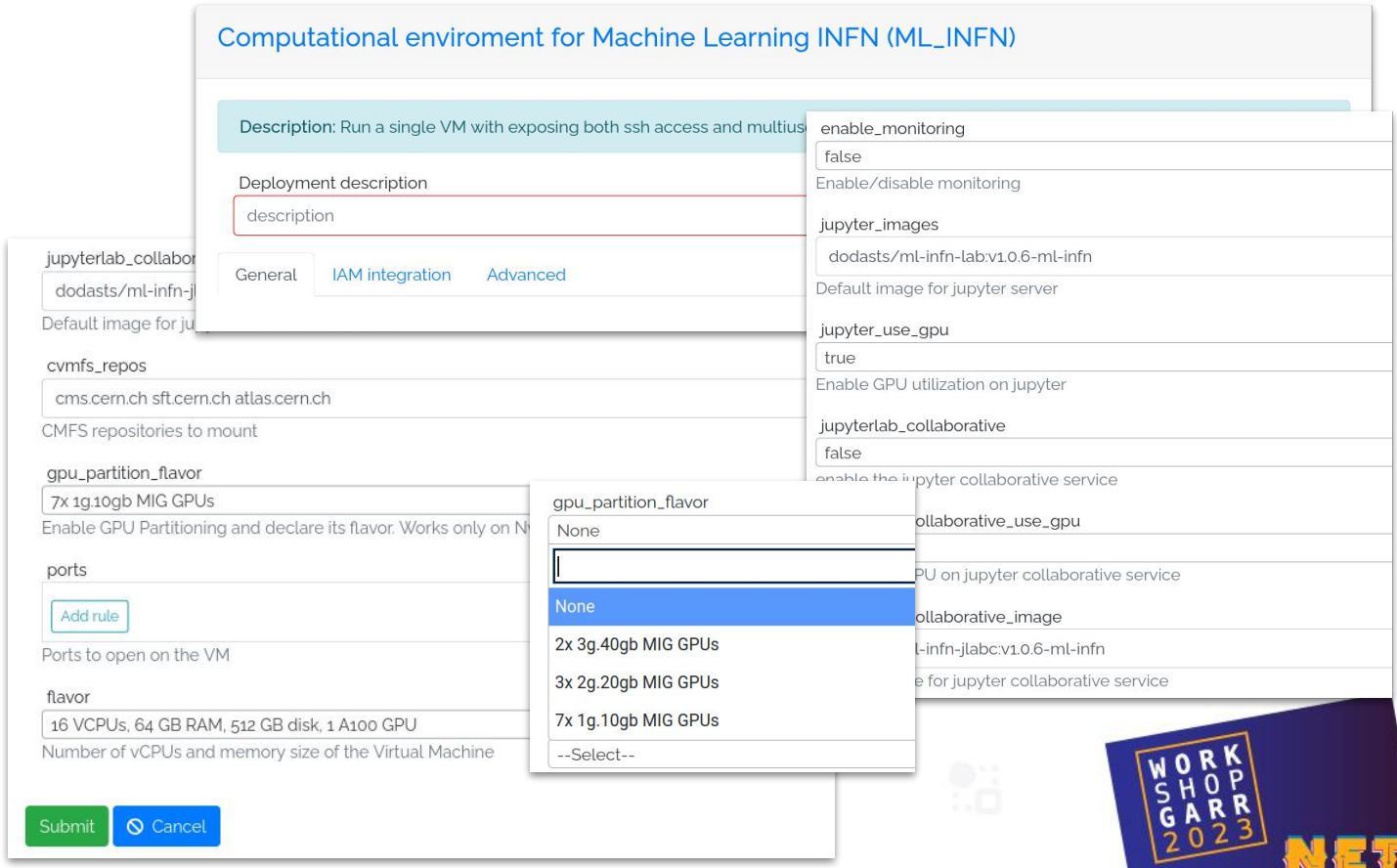
Phases:

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



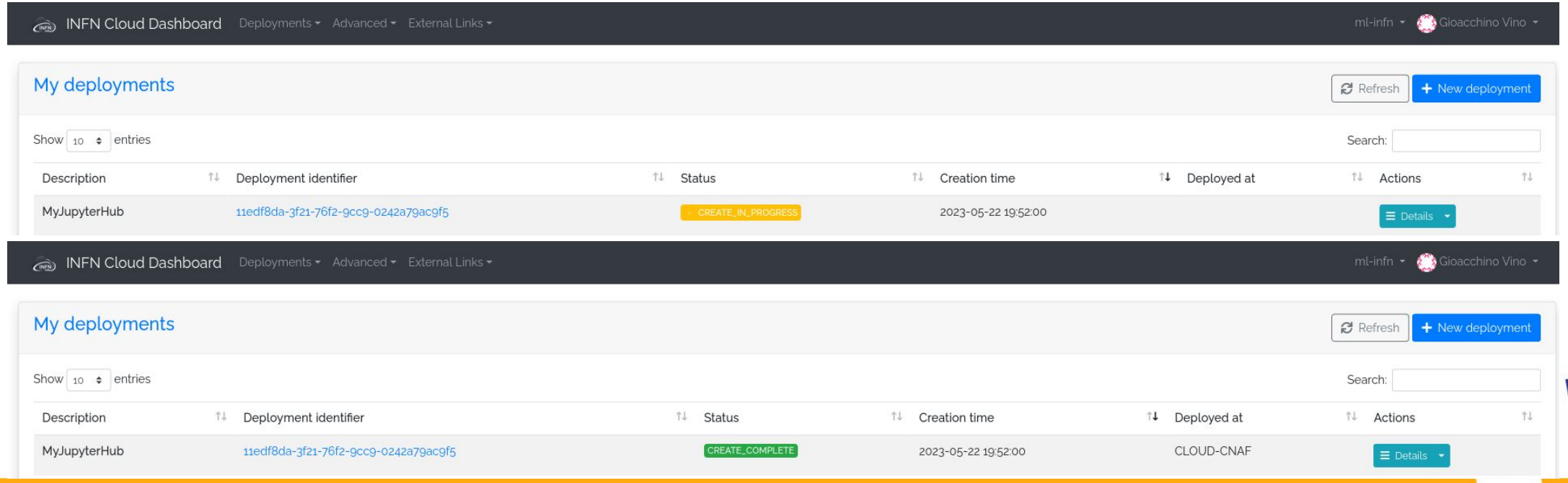
Computational environment for Machine Learning INFN

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



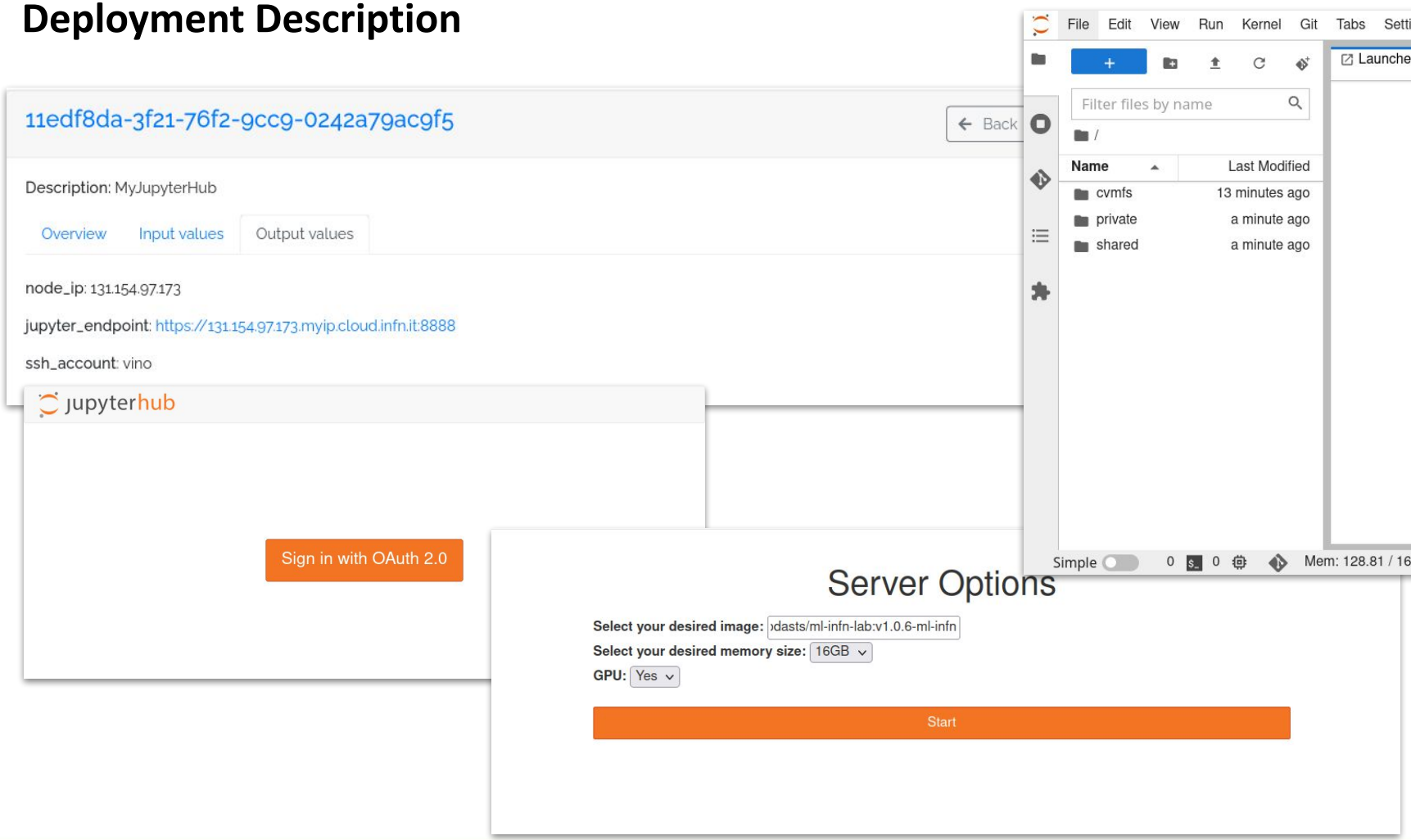
Computational environment for Machine Learning INFN

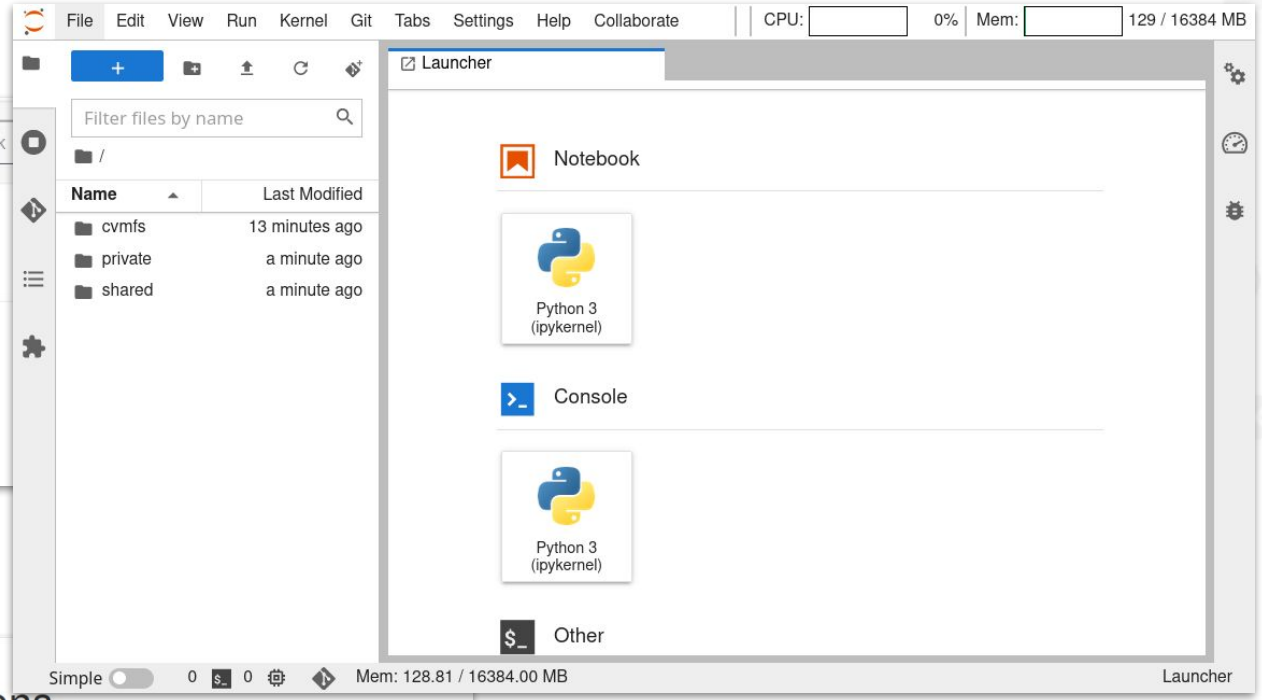
1. Collection of required information through the **INFN-Cloud Dashboard**
2. Evaluation if required **resource are available** (VM)
3. **Automatized configuration** on the service of the Virtual Machine



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

Contributors

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

Thanks for your attention

gioacchino.vino@infn.it

WORKSHOP GARR 2023

NET MAKERS

QUESTIONS?

www.garr.it/domande