

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



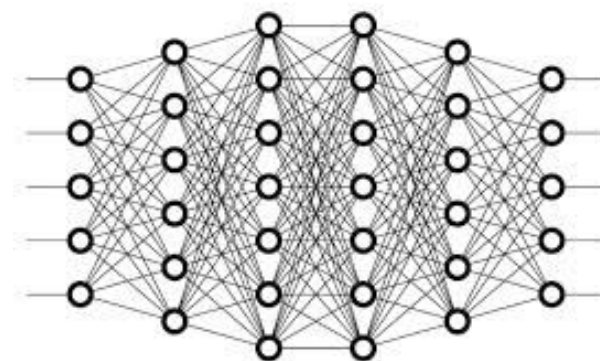
Gioacchino Vino
INFN

WORK
SHOP
GARR
2023

NET
MAKERS

Motivations

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



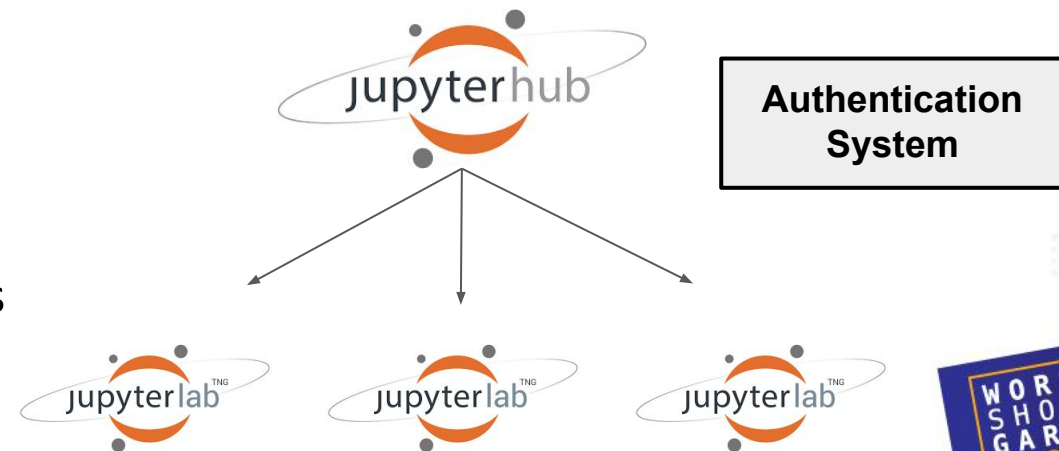
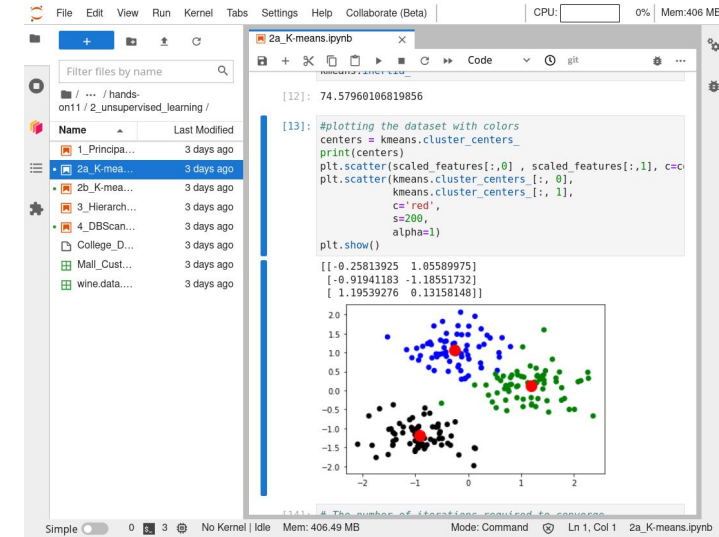
Jupyter

JupyterLab

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

JupyterHub

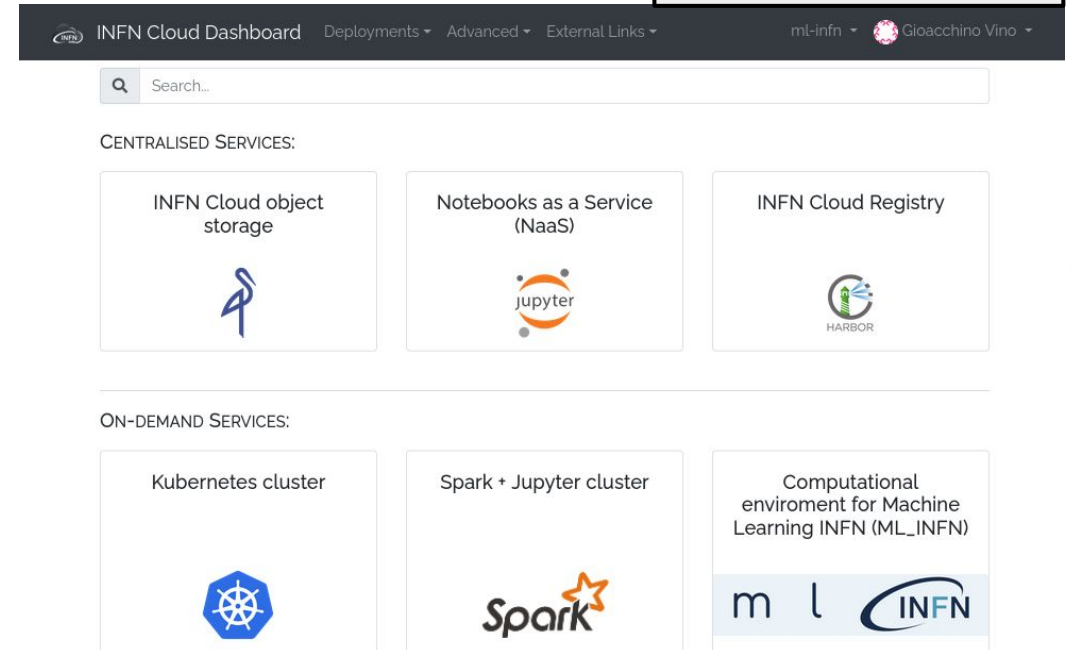
- Handles user authentication
- Extends JupyterLab to be a multi-user service
- Manages all accesses to different Jupyter instances
- Manages users and Jupyter instances



INFN-DataCloud Dashboard

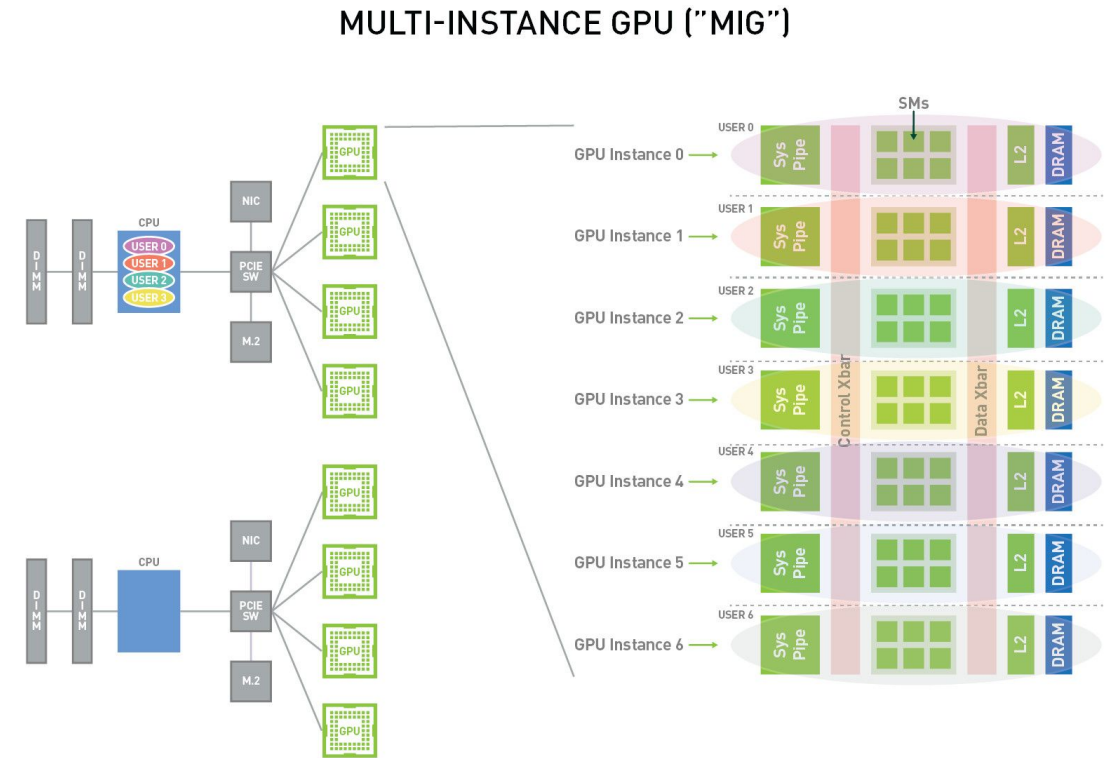
- Accessible to my.cloud.infn.it
- Central Portal
- Centralised Services and On-Demand Services
- Portfolio
 - Object Storage
 - Notebook as a Service
 - Registry
 - Kubernetes cluster
 - Spark + Jupyter
 - Computational environment for Machine Learning INFN

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



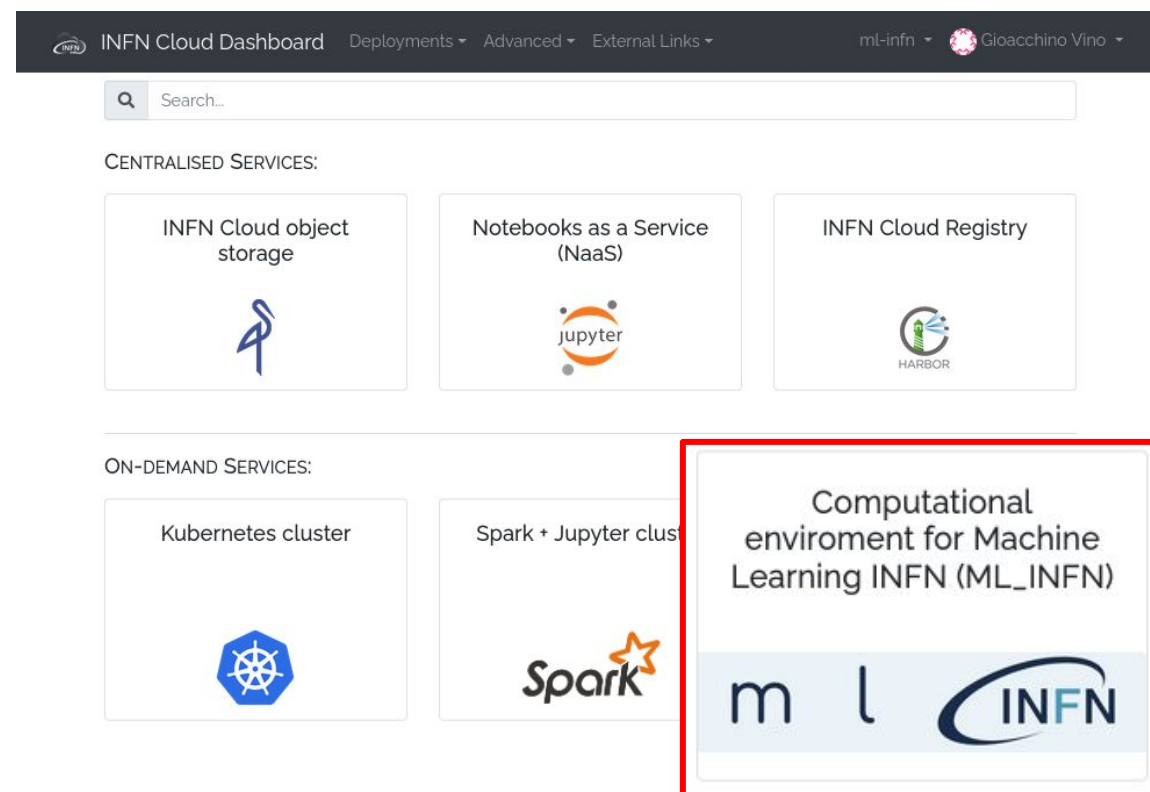
Computational environment for Machine Learning INFN

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

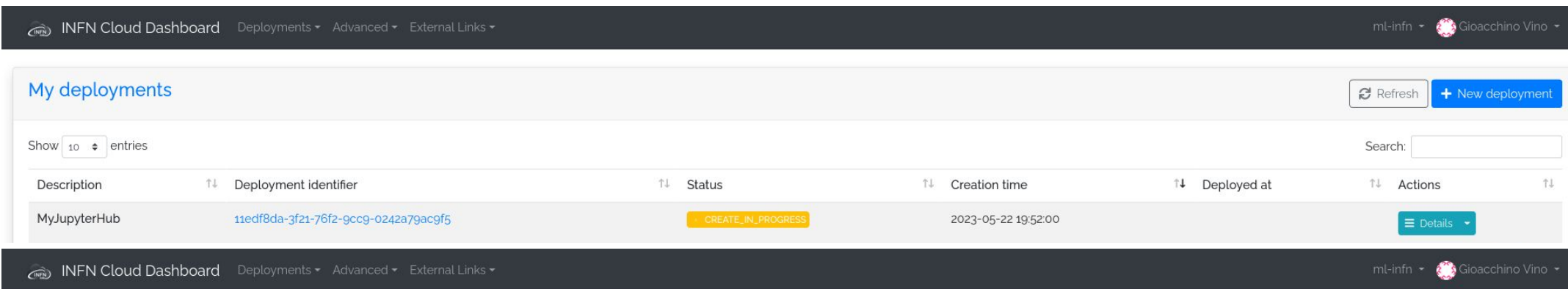
The screenshot displays the 'Computational environment for Machine Learning INFN (ML_INFEN)' configuration page. The page is divided into several sections:

- Description:** Run a single VM with exposing both ssh access and multiuser.
- Deployment description:** A text input field labeled 'description'.
- General tab:** Contains fields for 'jupyterlab_collaborative' (set to 'dodasts/ml-infn-jlab'), 'cvmfs_repos' (set to 'cms.cern.ch sft.cern.ch atlas.cern.ch'), 'gpu_partition_flavor' (set to '7x 1g.10gb MIG GPUs'), 'ports' (with an 'Add rule' button), and 'flavor' (set to '16 VCPUs, 64 GB RAM, 512 GB disk, 1 A100 GPU').
- IAM integration tab:** Contains a 'description' field.
- Advanced tab:** Contains a list of configuration options: 'enable_monitoring' (false), 'jupyter_images' (dodasts/ml-infn-lab:v1.0.6-ml-infn), 'jupyter_use_gpu' (true), 'jupyterlab_collaborative' (false), 'collaborative_use_gpu' (true), and 'collaborative_image' (ml-infn-jlab:v1.0.6-ml-infn).

A dropdown menu for 'gpu_partition_flavor' is open, showing options: 'None', '2x 3g.40gb MIG GPUs', '3x 2g.20gb MIG GPUs', '7x 1g.10gb MIG GPUs', and '--Select--'. The 'Submit' button is green, and the 'Cancel' button is blue.

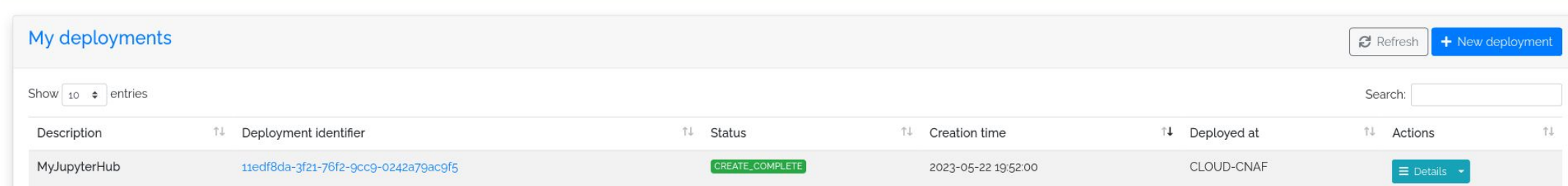
Computational environment for Machine Learning INFN

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



The screenshot shows the INFN Cloud Dashboard interface. The top navigation bar includes the INFN logo, "INFN Cloud Dashboard", and links for "Deployments", "Advanced", and "External Links". On the right, it shows the user "ml-infn" and "Gioacchino VINO". The main section is titled "My deployments" and includes a "Refresh" button and a "+ New deployment" button. Below this, there is a search bar and a table of deployments. The table has columns for Description, Deployment identifier, Status, Creation time, Deployed at, and Actions. A single deployment is listed with the description "MyJupyterHub", the identifier "11edf8da-3f21-76f2-9cc9-0242a79ac9f5", and a status of "CREATE_IN_PROGRESS". A "Details" button is visible next to the deployment.

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



This screenshot shows the same INFN Cloud Dashboard interface as the previous one, but the deployment status has changed to "CREATE_COMPLETE". The "Deployed at" field now shows "CLOUD-CNAF".

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

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



Sign in with OAuth 2.0

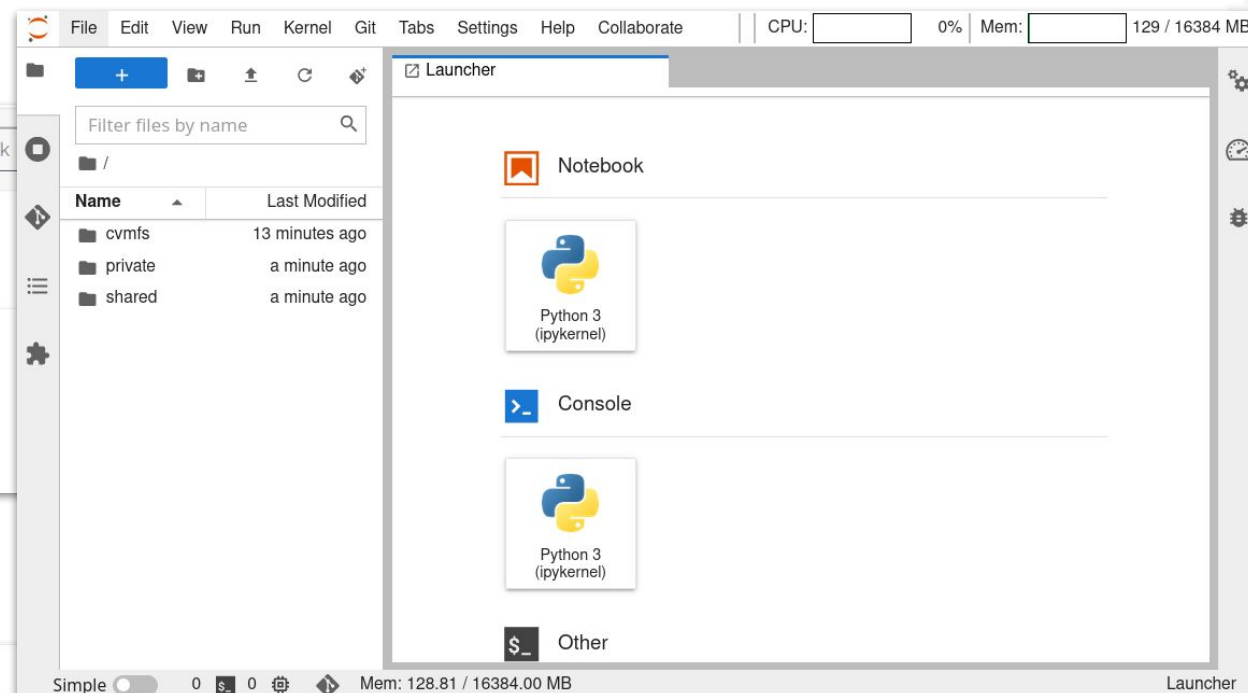
Server Options

Select your desired image:

Select your desired memory size:

GPU:

Start



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



**NET
MAKERS**

**Thanks for your
attention**

giacchino.vino@inf.n.it

QUESTIONS?



www.garr.it/domande