Present and future state of the platform (*user oriented*)

Andrés Heredia (heredia@ifca.unican.es)
Ignacio Heredia (iheredia@ifca.unican.es)

Co-funded by
the European Union

AI4EOSC Platform: User's workshop (November 2023)

AI4 | eosc

Present

Future

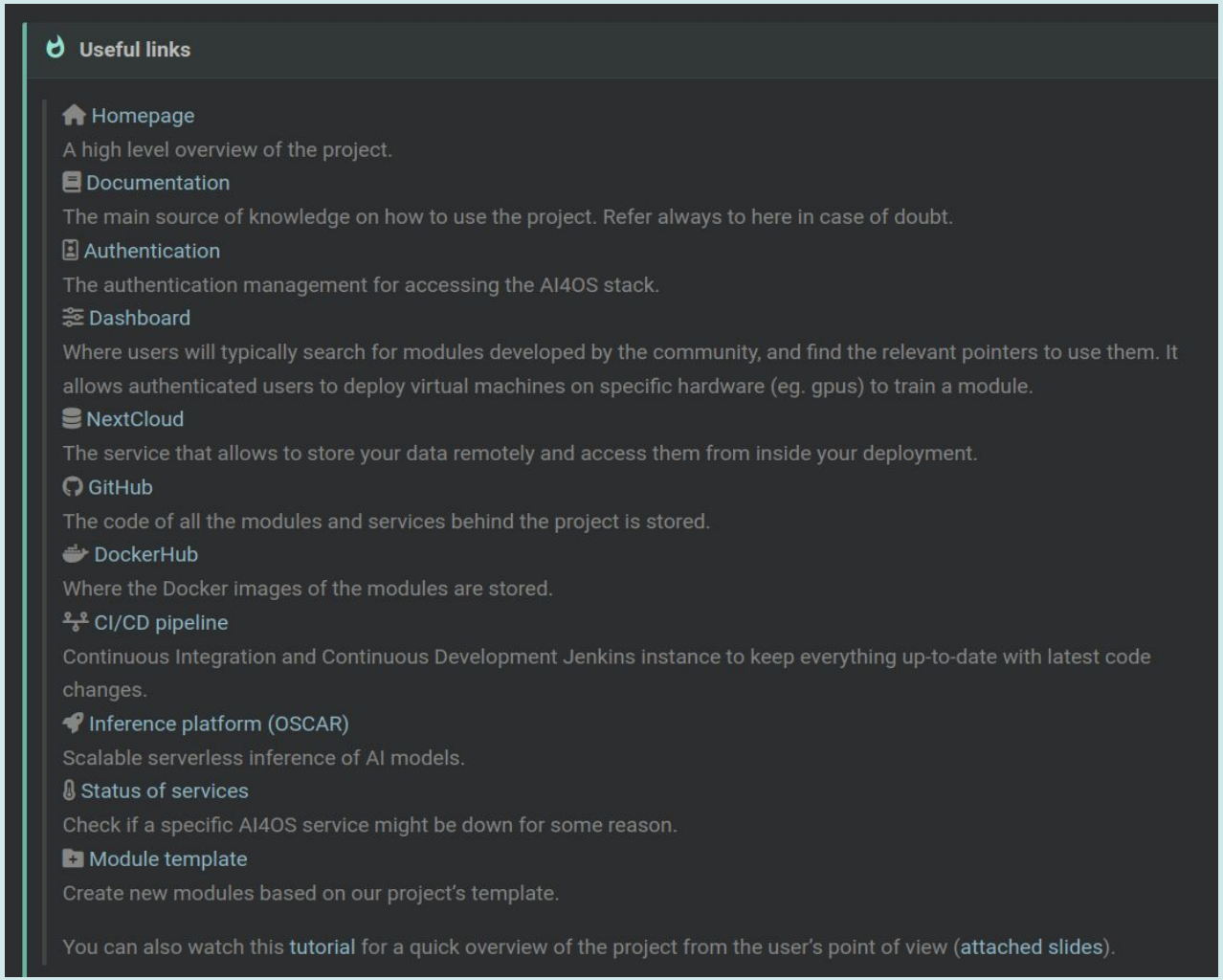
Co-funded by
the European Union

General overview - Docs

AI4 | eosc

Co-funded by the European Union

https://docs.ai4eosc.eu/



Useful links

Homepage
A high level overview of the project.
Documentation
The main source of knowledge on how to use the project. Refer always to here in case of doubt.
Authentication
The authentication management for accessing the AI4OS stack.
Dashboard
Where users will typically search for modules developed by the community, and find the relevant pointers to use them. It allows authenticated users to deploy virtual machines on specific hardware (eg. gpus) to train a module.
NextCloud
The service that allows to store your data remotely and access them from inside your deployment.
GitHub
The code of all the modules and services behind the project is stored.
DockerHub
Where the Docker images of the modules are stored.
CI/CD pipeline
Continuous Integration and Continuous Development Jenkins instance to keep everything up-to-date with latest code changes.
Inference platform (OSCAR)
Scalable serverless inference of AI models.
Status of services
Check if a specific AI4OS service might be down for some reason.
Module template
Create new modules based on our project's template.

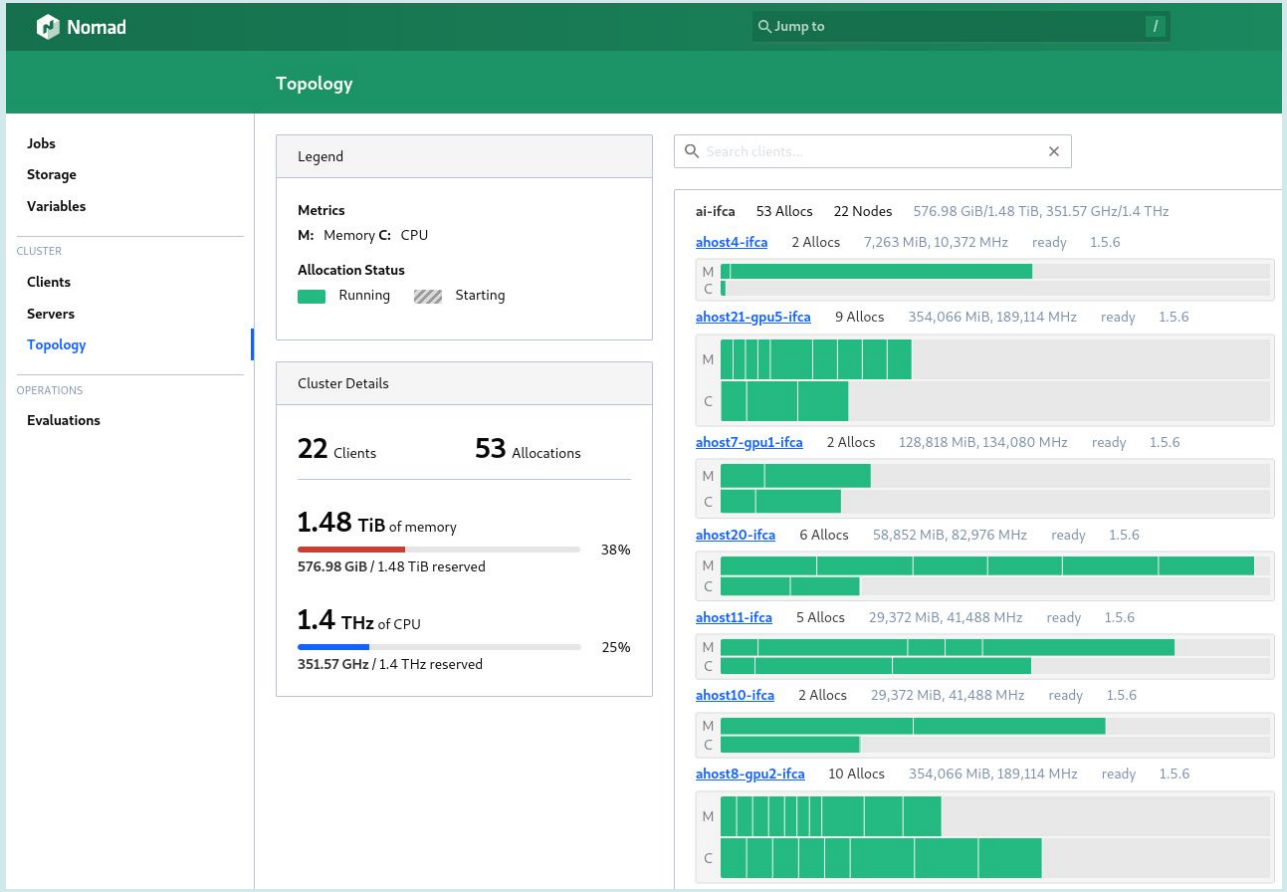
You can also watch this tutorial for a quick overview of the project from the user's point of view (attached slides).

The Nomad cluster

AI4 | eosc

Co-funded by the European Union

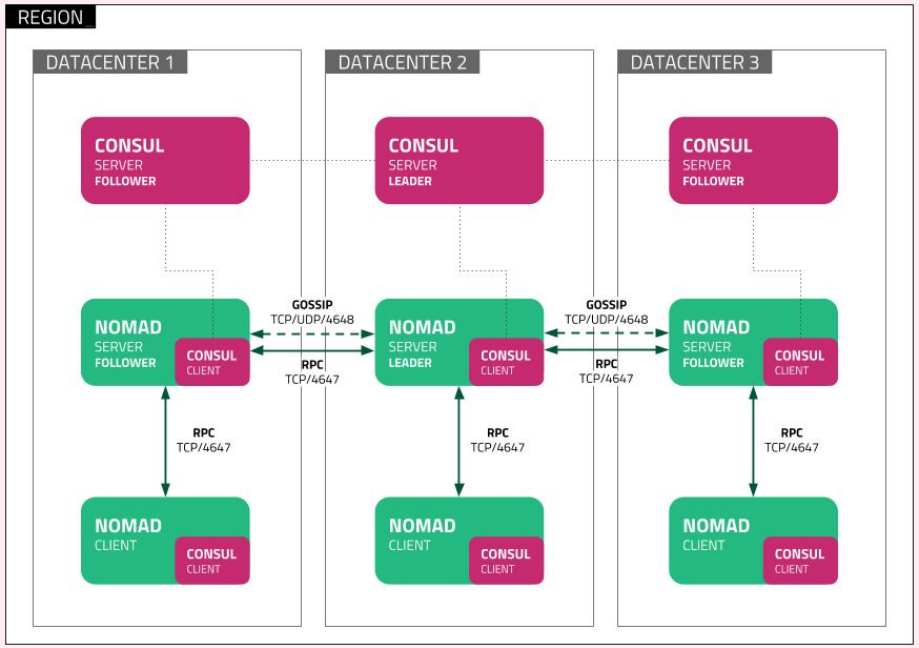
- one production cluster (IFCA) with 25 GPUs.
- multiple development clusters deployed for different tests (distributed/federated, full service mesh).
- deployments can make use of the “/storage” paths to automatically sync with Nextcloud (! slower access to files though)

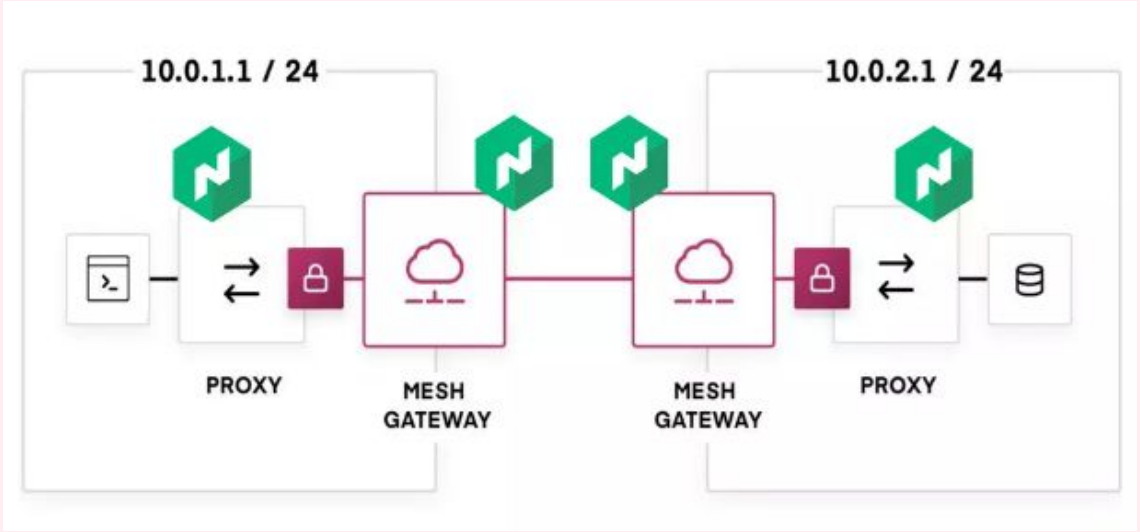


The Nomad cluster

AI4 | eosc
Co-funded by the European Union

- add more resources (GPUs, storage, …)
- properly limit storage space inside each deployment
- improve data access (share volumes between deployments to avoid copying data from Nextcloud)
- federate multiple providers (IISAS, LIP, …)





The Nomad cluster

AI4 | eosc

Co-funded by the European Union

⚠️ **November 27th** yearly maintenance shutdown

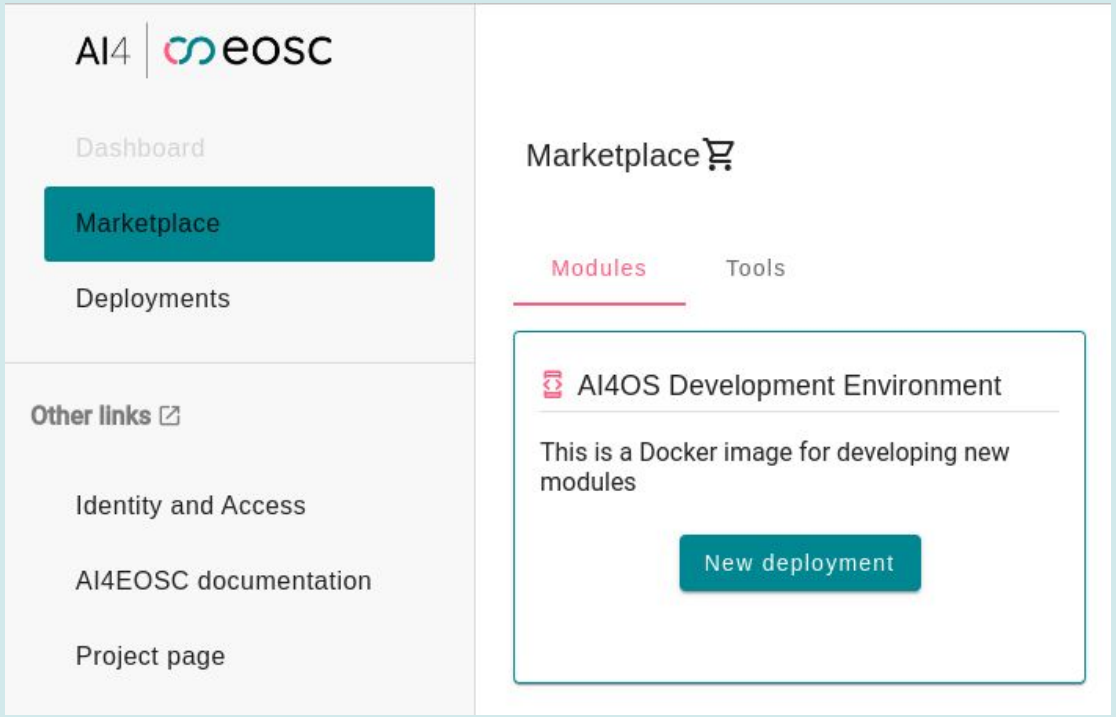
Please, don’t forget to do a **backup** of your running jobs!

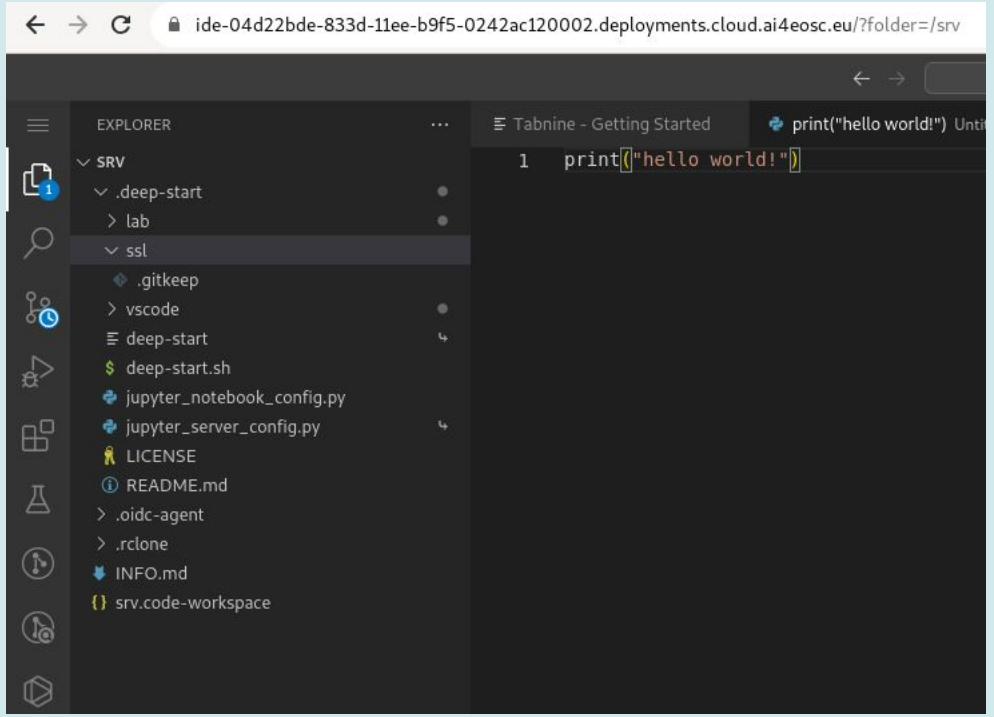
PAPI + Dashboard

AI4 | eosc

Co-funded by the European Union

- Tools/Deployments view
- SSL in deployments
- Dev module with main DL frameworks and VS code installed
- Federated server tool (flower.io) available to perform federated trainings





PAPI + Dashboard

AI4 | eosc Co-funded by the European Union

- Experiment centric view:
  - 1 experiment gathers several deployments and tools
  - ability to share experiments across users
- Visualize cluster stats in the Dashboard:
  - free resources
  - current user usage
  - historic user usage
  - historic VO usage
- Filter modules by tags
- New tools to be added:
  - MLflow (experiment tracking)
  - CVAT (image annotation)
- Deploy services specifically tailored for inference (ie. automatic scaling) (using OSCAR under the hood)
- Authentication for clients in federated trainings via token

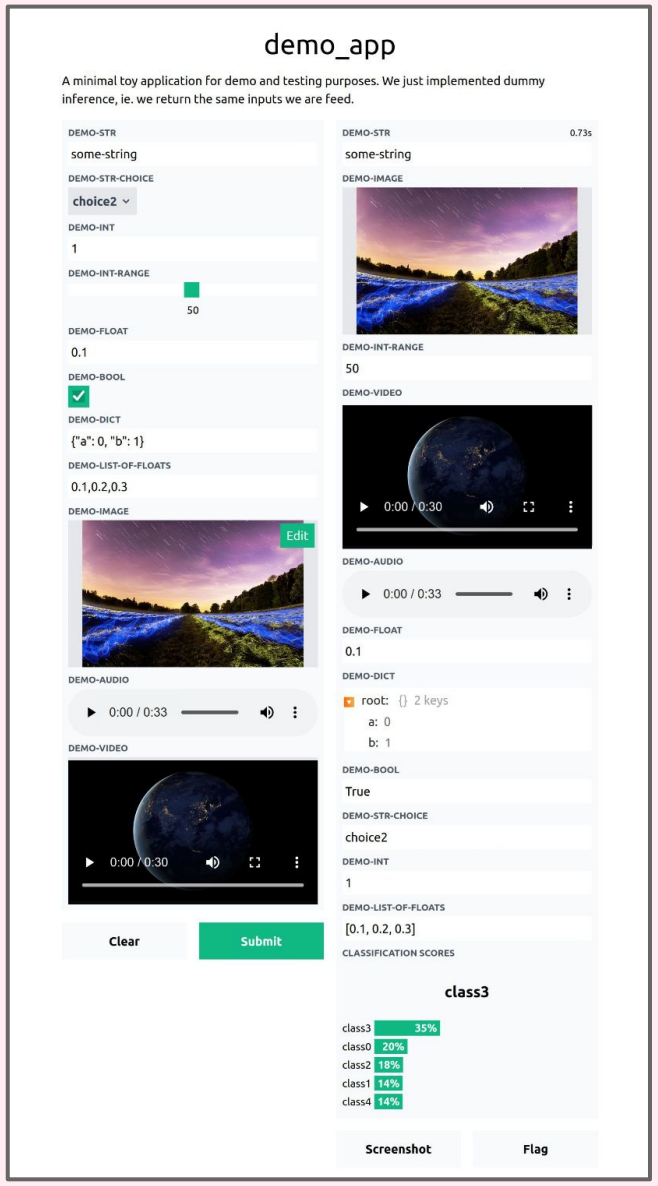
Modules

AI4 | eosc Co-funded by the European Union

- Module creation with AI4OS template (cookiecutter)
- Module metadata in JSON
- DEEPaaS API to expose functionality of the module

Modules

- (?) Unify repos
- (?) Easier Jenkinsfiles
- Metadata:
  - move to YAML format
  - more tag categories
  - remove clutter
- DEEPaaS:
  - (?) type hints to define input args
  - (?) decorator use decorator instead of entrypoints
  - Integrate Gradio UI inside main DEEPaaS for inference endpoints



Others

- Improved CI/CD management (dedicated machine)
- Project auth:
  - Migration to EGI production auth (iMagine users)
  - (?) Migration to IAM (AI4EOSC users)
- Migration to new Nextcloud instance (hopefully transparent)
- Compose multistep inference flows with OSCAR, Elyra and Node-red.

What do ***you*** want to see in the platform?

feature requests - rough edges - enhancements - …

AI4 | eosc

Co-funded by
the European Union