ECAS: a data science environment for climate change in the EGI federated infrastructure

Fabrizio Antonio, Donatello Elia, Sandro Fiore

On behalf of the ECAS Team

eosc-hub.eu

@EOSC_eu

EOSC-hub week 2020

Online | 18-20 May 2020

Impact of EOSC-hub on science communities

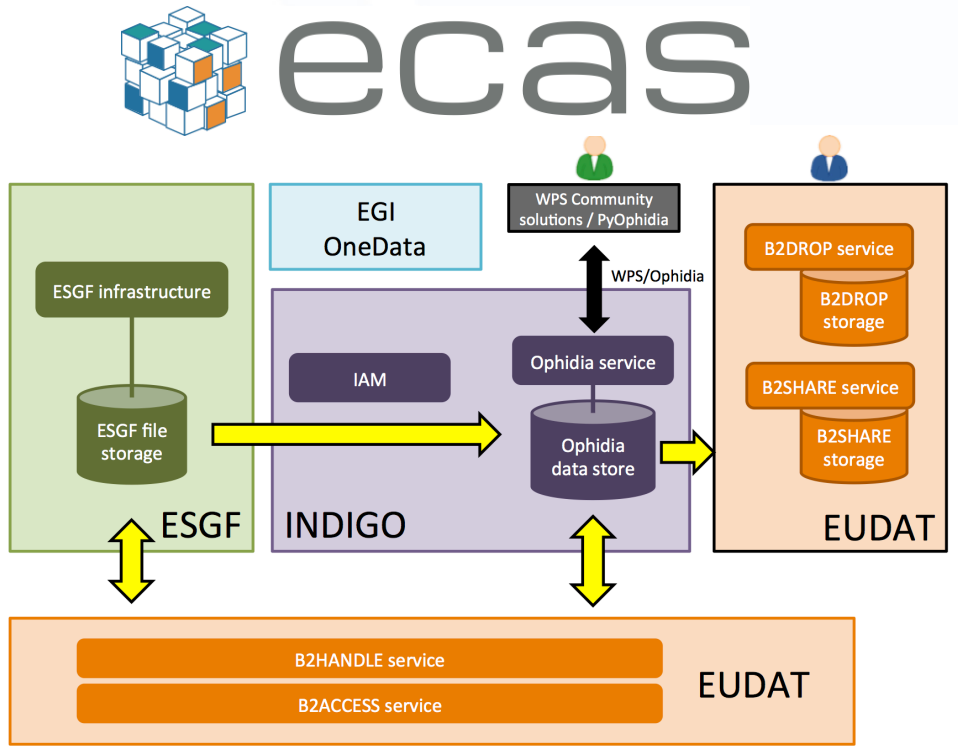
EOSC-hub receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 777536.

ENES Climate Analytics Service (ECAS)

The *ENES Climate Analytics Service (**ECAS**)*, proposed by CMCC & DKRZ in the *EU H2020 EOSC-Hub project*, supports **climate data analysis experiments** with a strong focus on **data intensive analysis**, **provenance management**, and **server-side approaches**

It is one of the ***EOSC-Hub Thematic Services*** *as well as* ***a Compute Service*** *in the* ***IS-ENES3 project***

ECAS consists of multiple integrated components from INDIGO-DataCloud, EUDAT, ESGF and EGI, centered around the ***Ophidia HPDA framework***



Ophidia

is-enes
INFRASTRUCTURE FOR THE EUROPEAN NETWORK FOR EARTH SYSTEM MODELLING

EOSC-hub

https://www.eosc-hub.eu/services/ENES%20Climate%20Analytics%20Service

ECASLab: a Python environment for data analysis

ECASLab provides a user-friendly environment for scientific analysis based on:

The *ECAS* integrated service

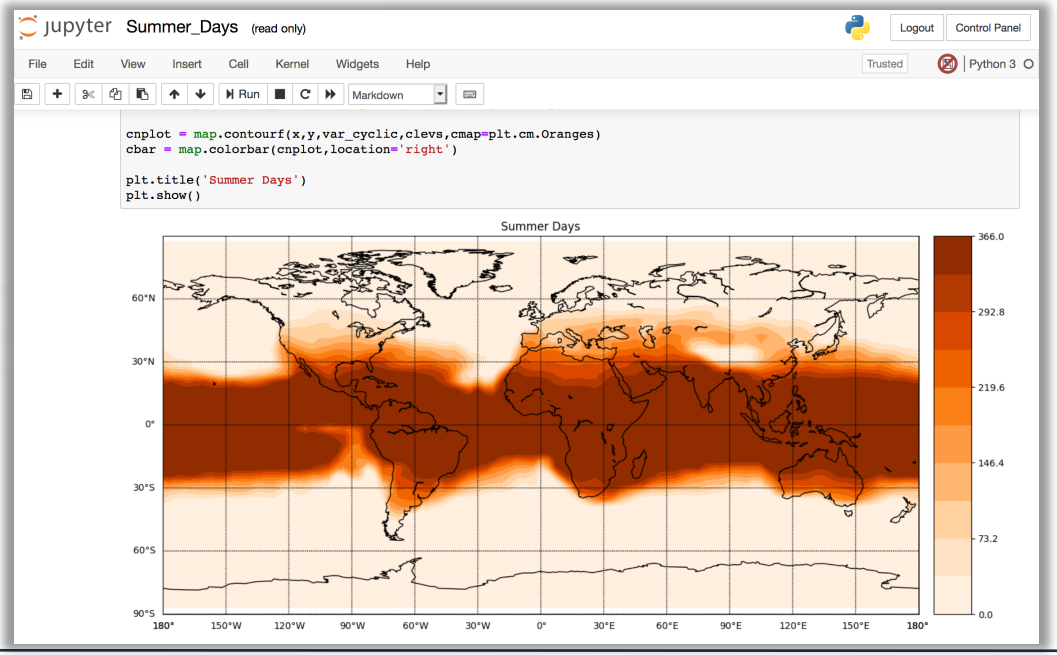
A ***JupyterHub*** instance providing a graphical environment for users experiments

A wide set of ***Python scientific modules*** for data manipulation, analysis and visualization (PyOphidia, NumPy, Pandas, Dask, Matplotlib, basemap, Cartopy)

A set of ECAS usage example **notebooks** (https://github.com/ECAS-Lab/ecas-notebooks)

Two major instances hosted by:

- ✓ **CMCC** https://ecaslab.cmcc.it
- ✓ **DKRZ** https://ecaslab.dkrz.de



The Ophidia project

Ophidia (http://ophidia.cmcc.it) is a CMCC Foundation research project addressing data challenges for eScience[1]

It provides:

a *High Performance Data Analytics* (**HPDA**) framework for multi-dimensional scientific data joining HPC paradigms with scientific data analytics approaches

in-memory and **server-side** data analysis exploiting parallel computing techniques and database approaches

a **multi-dimensional**, **array-based**, storage model and partitioning schema for scientific data leveraging the datacube abstraction

end-to-end mechanisms to support **complex experiments and large workflows** on scientific datacubes, primarily in **climate domain**

Ophidia

cmcc
Centro Euro-Mediterraneo sui Cambiamenti Climatici

1. S. Fiore, A. D'Anca, C. Palazzo, I. T. Foster, D. N. Williams, G. Aloisio, "Ophidia: toward big data analytics for escience", *ICCS 2013*

Programmatic access through the PyOphidia class

PyOphidia provides a Python interface to submit commands to the Ophidia Server and to easily retrieve the results (e.g. in Jupyter Notebooks)

Two modules available:

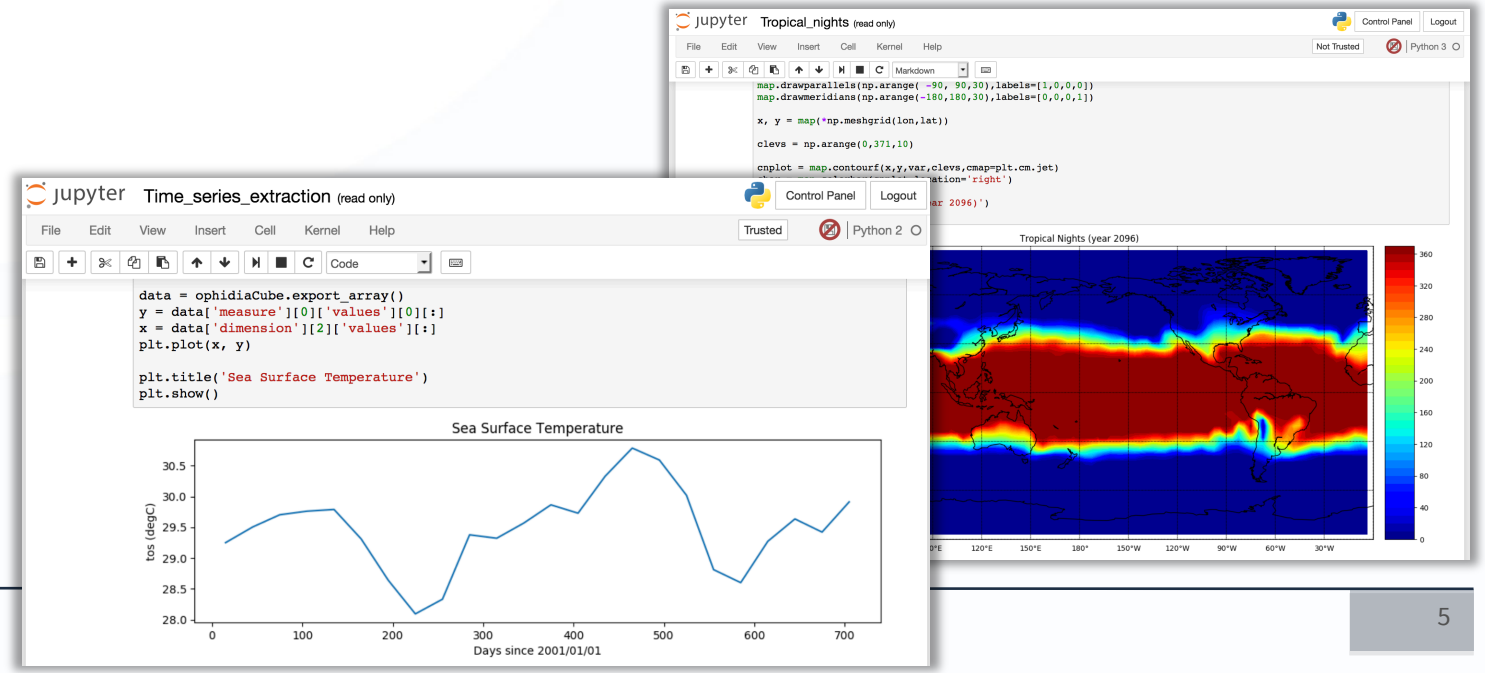
Client *class*: supports the submissions of Ophidia commands and workflows as well as the management of sessions

Cube *class*: provides the datacube type abstraction and the methods to manipulate, process and get information from cubes objects

```
from PyOphidia import cube, client
cube.Cube.setclient(read_env=True)

mycube =
cube.Cube.importnc(src_path='/public/data/ecas_training
/file.nc', measure='tos', imp_dim='time',
import_metadata='yes', ncores=5)
mycube2 = mycube.reduce(operation='max',ncores=5)
mycube3 = mycube2.rollup(ncores=5)
data = mycube3.export_array()

mycube3.exportnc2(output_path='/home/test',
export_metadata='yes')
```



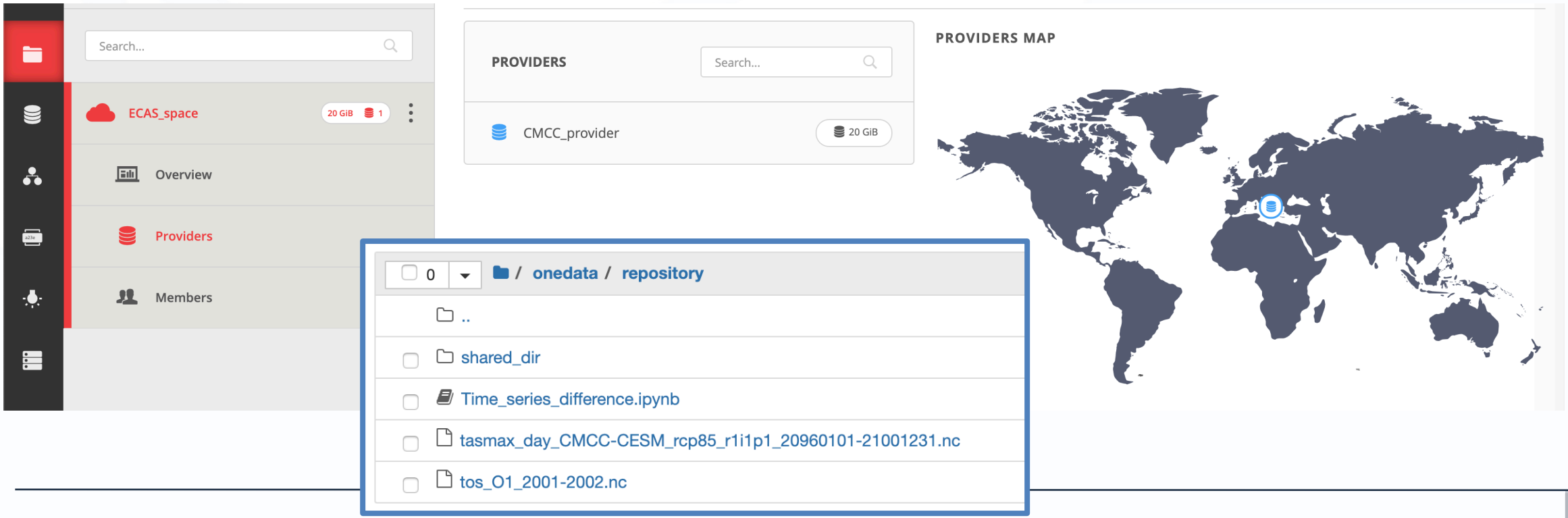
https://github.com/OphidiaBigData/PyOphidia
https://pypi.org/project/PyOphidia/
https://anaconda.org/conda-forge/pyophidia

Integration of ECAS with Onedata

Onedata Provider *deployed at CMCC SuperComputing Center to support a shared read-only repository (**ECAS_space**)*

*Single **OneClient** instance set up to interact with the provider*

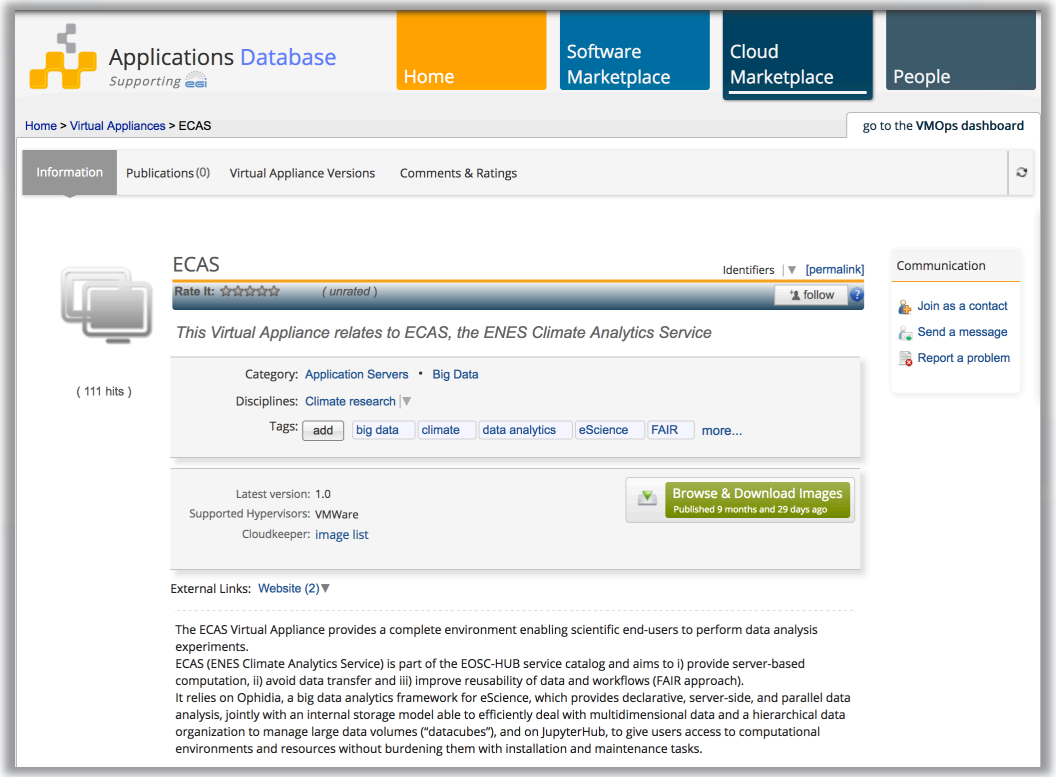
*Data folders mounted on the ECAS users home through **NFS** in read-only mode*



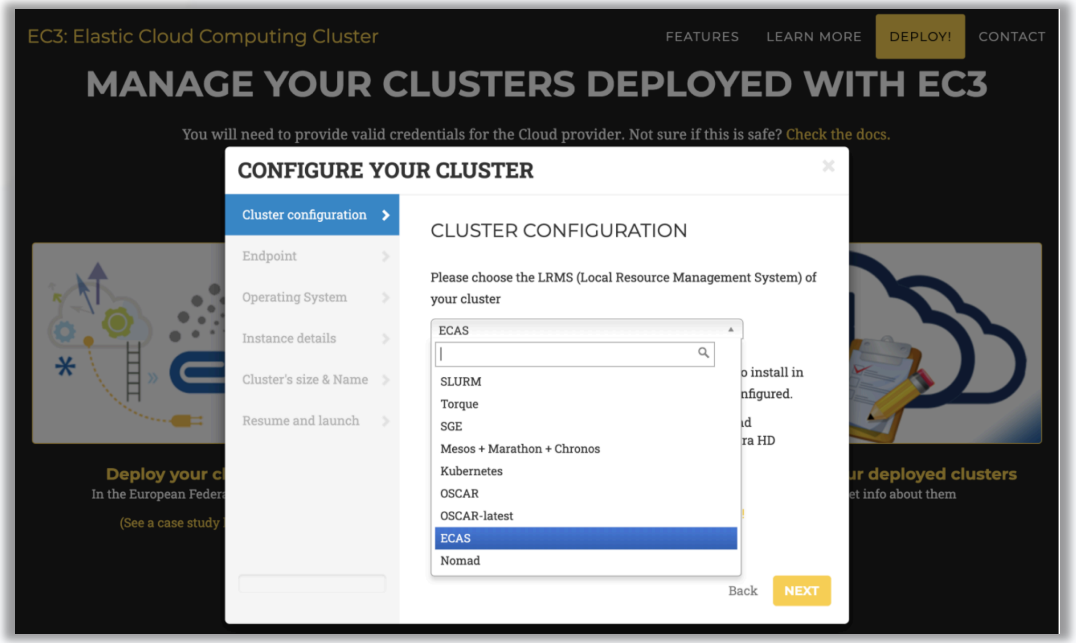
ECAS on the EGI Federated Cloud Infrastructure

ECAS has been integrated into the EGI FedCloud, considering two scenarios[1]:

A ready-to-use ECAS single-node VMI available from the EGI AppDB https://appdb.egi.eu/store/vappliance/ecas



A multi-node ECAS cluster dinamically provisioned on the EGI FedCloud through the Elastic Cloud Computing Cluster (EC3)

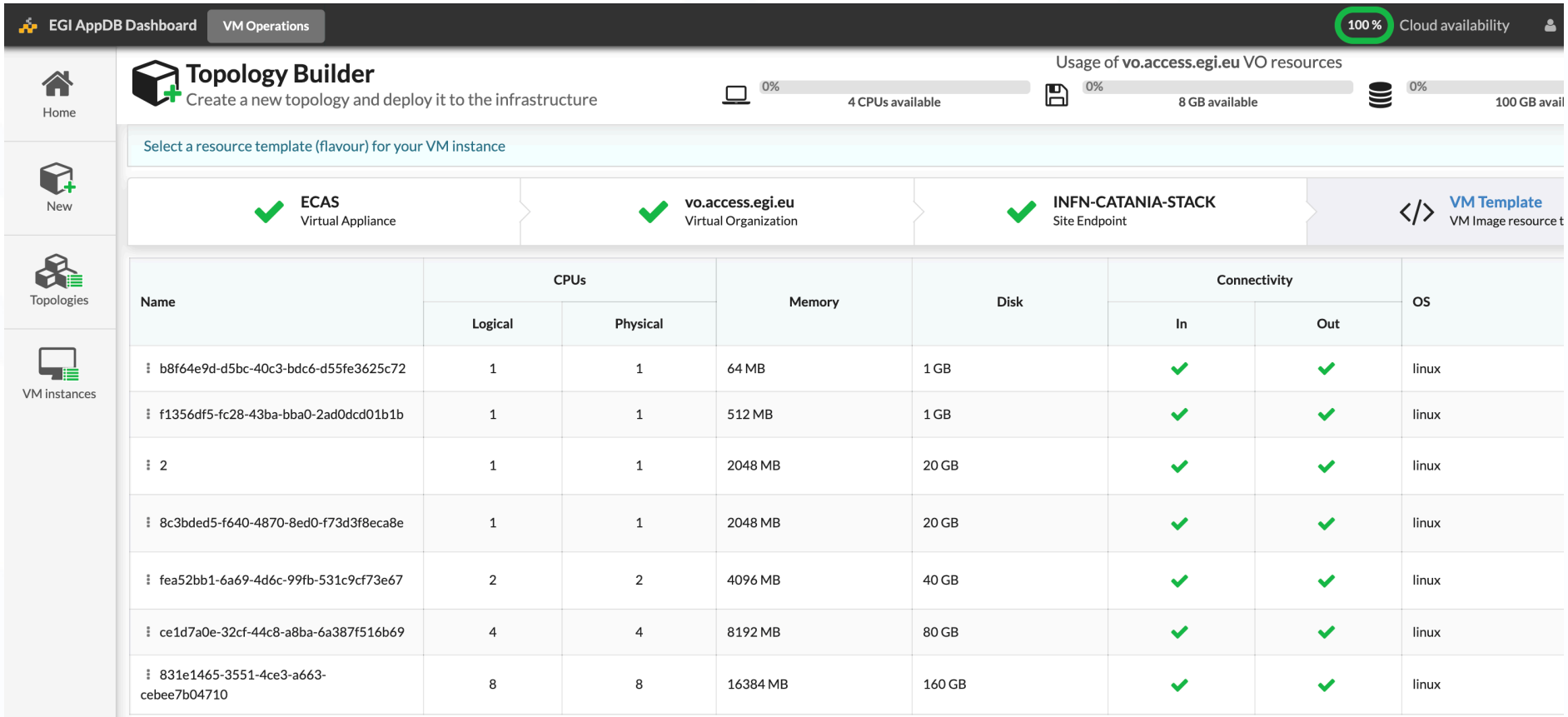


1. Elastic deployment of ECAS on EGI: https://www.egi.eu/about/newsletters/elastic-deployment-of-ecas-on-egi/

ECAS single-node VMI

ECAS single-instance VMI *uploaded to the* ***EGI AppDB***

VMI *assigned to a set of* ***trusted VOs***



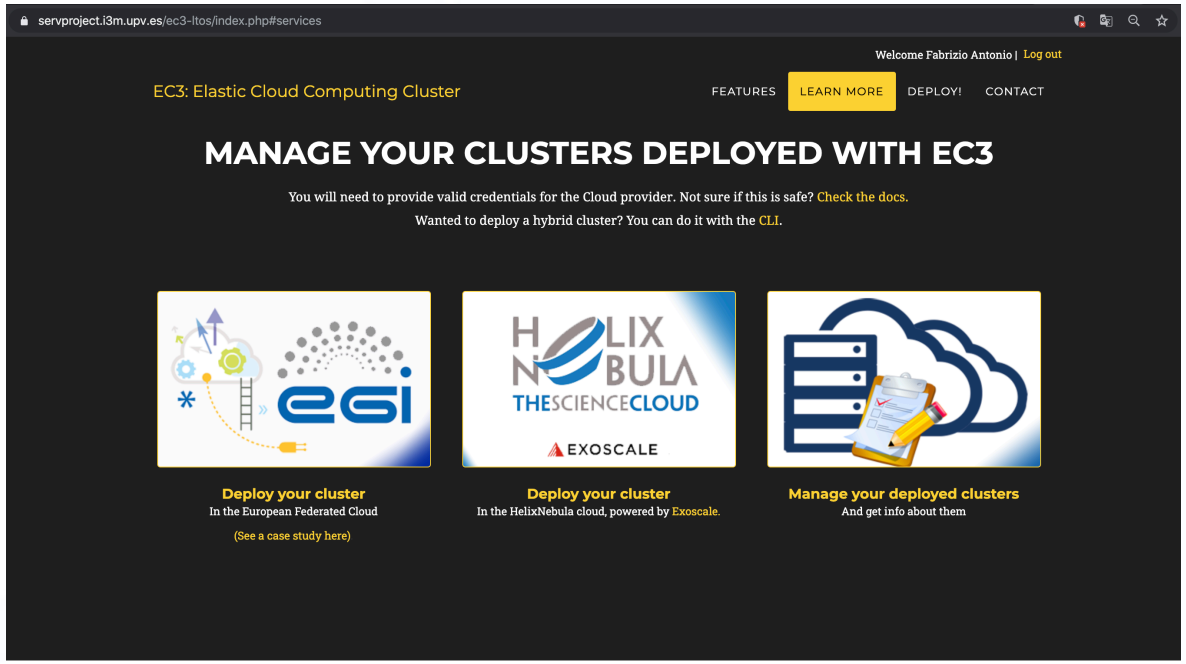
Name	CPUs		Memory	Disk	Connectivity		OS
	Logical	Physical			In	Out	
b8f64e9d-d5bc-40c3-bdc6-d55fe3625c72	1	1	64 MB	1 GB	✓	✓	linux
f1356df5-fc28-43ba-bba0-2ad0dcd01b1b	1	1	512 MB	1 GB	✓	✓	linux
2	1	1	2048 MB	20 GB	✓	✓	linux
8c3bded5-f640-4870-8ed0-f73d3f8eca8e	1	1	2048 MB	20 GB	✓	✓	linux
fea52bb1-6a69-4d6c-99fb-531c9cf73e67	2	2	4096 MB	40 GB	✓	✓	linux
ce1d7a0e-32cf-44c8-a8ba-6a387f516b69	4	4	8192 MB	80 GB	✓	✓	linux
831e1465-3551-4ce3-a663-cebee7b04710	8	8	16384 MB	160 GB	✓	✓	linux

Elastic deployment of ECAS on EGI

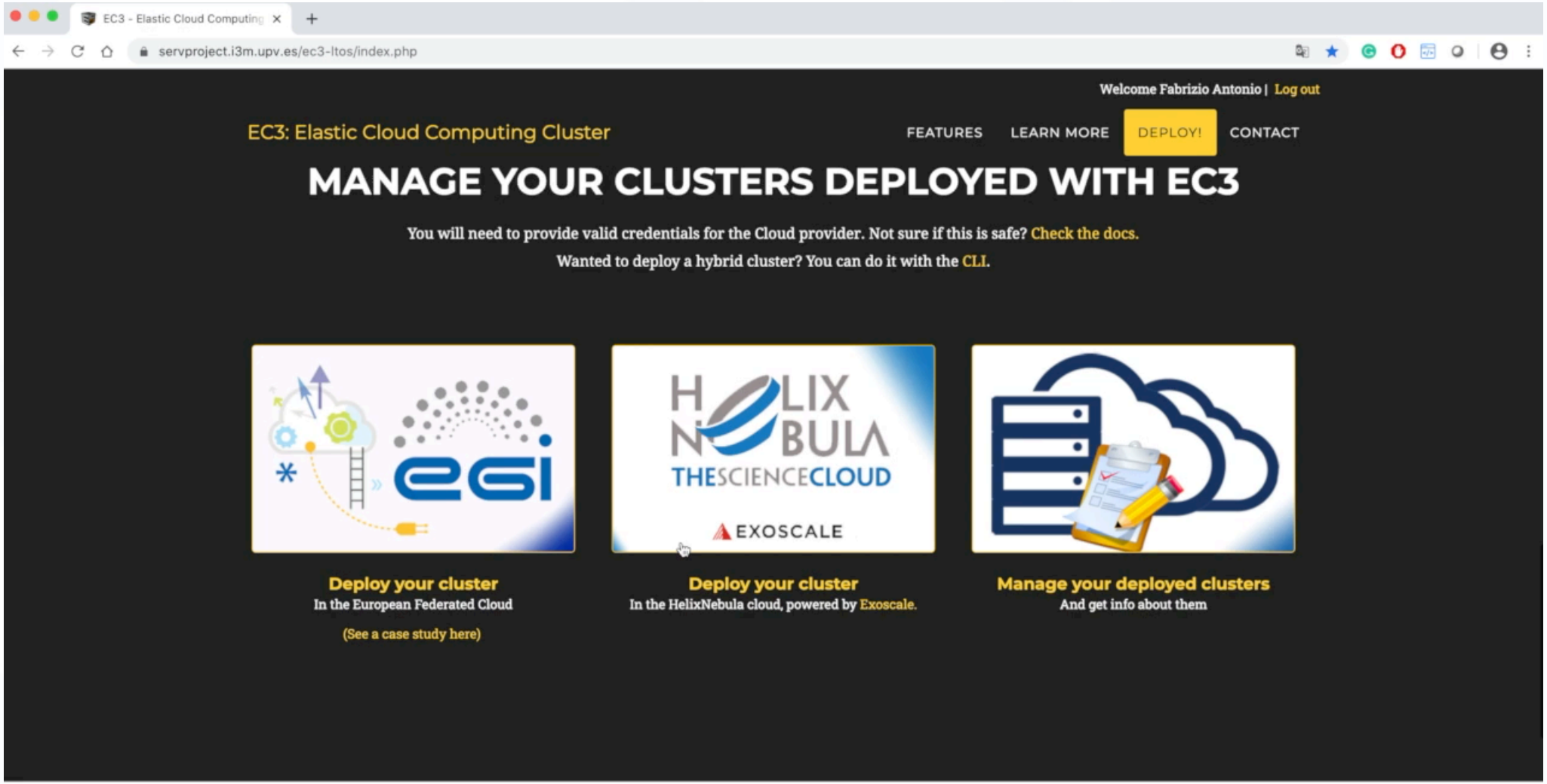
Multi-node ECAS environment *dynamically provisioned on the* ***FedCloud*** *through the* ***EC3*** ***LToS*** *service according to the user requirements*

RADL file *for the Infrastructure Manager (****IM****) to define the cluster setup in terms of resources, infrastructure and software configuration, and contextualization*

DEMO



DEMO


EC3 - Elastic Cloud Computing
servproject.i3m.upv.es/ec3-ltos/index.php
Welcome Fabrizio Antonio | Log out
EC3: Elastic Cloud Computing Cluster
FEATURES
LEARN MORE
DEPLOY!
CONTACT
MANAGE YOUR CLUSTERS DEPLOYED WITH EC3
You will need to provide valid credentials for the Cloud provider. Not sure if this is safe? Check the docs.
Wanted to deploy a hybrid cluster? You can do it with the CLI.
egi
HELIX NEBULA
THESCIENCECLOUD
EXOSCALE
Deploy your cluster
In the European Federated Cloud
(See a case study here)
Deploy your cluster
In the HelixNebula cloud, powered by Exoscale.
Manage your deployed clusters
And get info about them

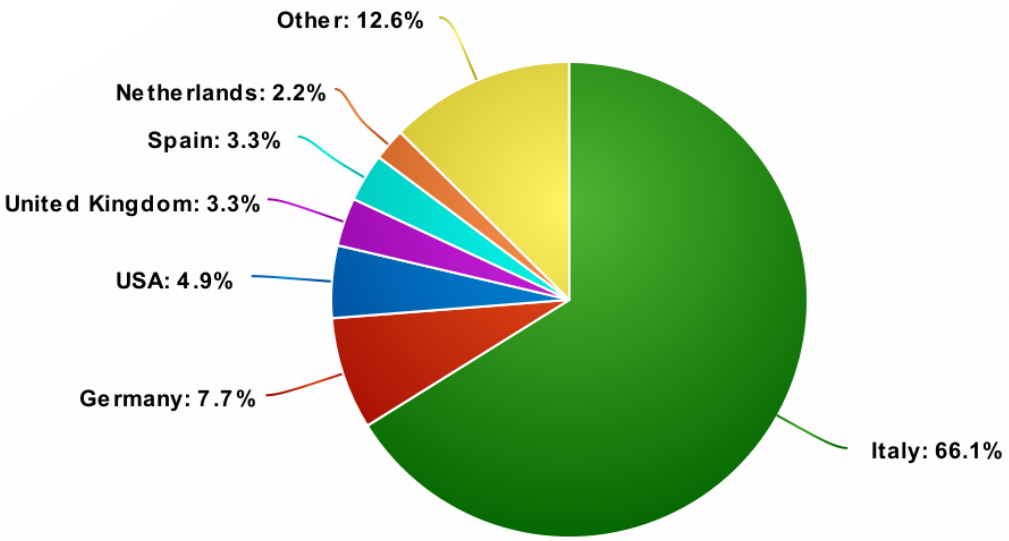
ECAS accounting management

timestamp	idworkflow	name	username	ip_address	client_address	#tasks	#success_tasks	duration
2018-03-29 01:08:48	1171802	oph_list	*4E6FA15F82F3...	192.168.88.150	193.204.199.213	1	1	0.194882
2018-03-29 01:09:05	1171804	oph_randcube	*4E6FA15F82F3...	192.168.88.150	193.204.199.213	1	1	0.613289
2018-03-29 08:29:54	1171806	oph_list	*0C0EBF0FBAAF...	192.168.88.150	193.204.199.213	1	1	0.185737
2018-03-29 09:16:38	1171808	oph_list	*7041(					
2018-03-29 09:36:34	1171810	oph_list	*7041(					

timestamp	idtask	idworkflow	operator	#cores	success_flag	duration
2018-03-29 01:08:48	1171803	1171802	oph_list	1	1	0.153245
2018-03-29 01:09:05	1171805	1171804	oph_randcube	1	1	0.578767
2018-03-29 08:29:54	1171807	1171806	oph_list	1	1	0.152150
2018-03-29 09:16:38	1171809	1171808	oph_list	1	1	0.223756
2018-03-29 09:36:34	1171811	1171810	oph_list	1	0	0.153697

Ophidia Server logs

Registered users **180**



Number of jobs **1750K**

Number of core-hours **~10K**

Other metrics about users and computing resources usage

https://github.com/ECAS-Lab/ecas-accounting

Training events **8**

Conclusions & next steps

Conclusions

ECAS enables scientific end-users to perform data analysis experiments on large volumes of multidimensional data by exploiting a PID-enabled, server-side, and parallel approach

The integration of **ECAS** in the **EGI cloud-based resources** allows researchers to deploy on demand a full ECAS elastic cluster on the **EGI Infrastructure** thanks to the **EC3 platform**.

Next steps

Integration of **EGI Check-in** in ECAS

Stronger integration with **Onedata**

Useful links

CMCC ECASLab instance: https://ecaslab.cmcc.it/

DKRZ ECASLab instance: https://ecaslab.dkrz.de/

ECASLab repository: https://github.com/ECAS-Lab

Ophidia Website: http://ophidia.cmcc.it

Ophidia Doc: http://ophidia.cmcc.it/documentation

PyOphidia repository: https://github.com/OphidiaBigData/PyOphidia

Acknowledgment

These activities are supported in part by EOC-Hub and IS-ENES3 projects:

EOSC-hub receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 777536

IS-ENES3 is a project funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 824084

EOSC-hub

Thanks