

EuroScienceGateway

Leveraging the European compute infrastructures for data-intensive research guided by FAIR principles



**Funded by
the European Union**
Grant agreement 101057388

2024-10-02 by EuroScienceGateway



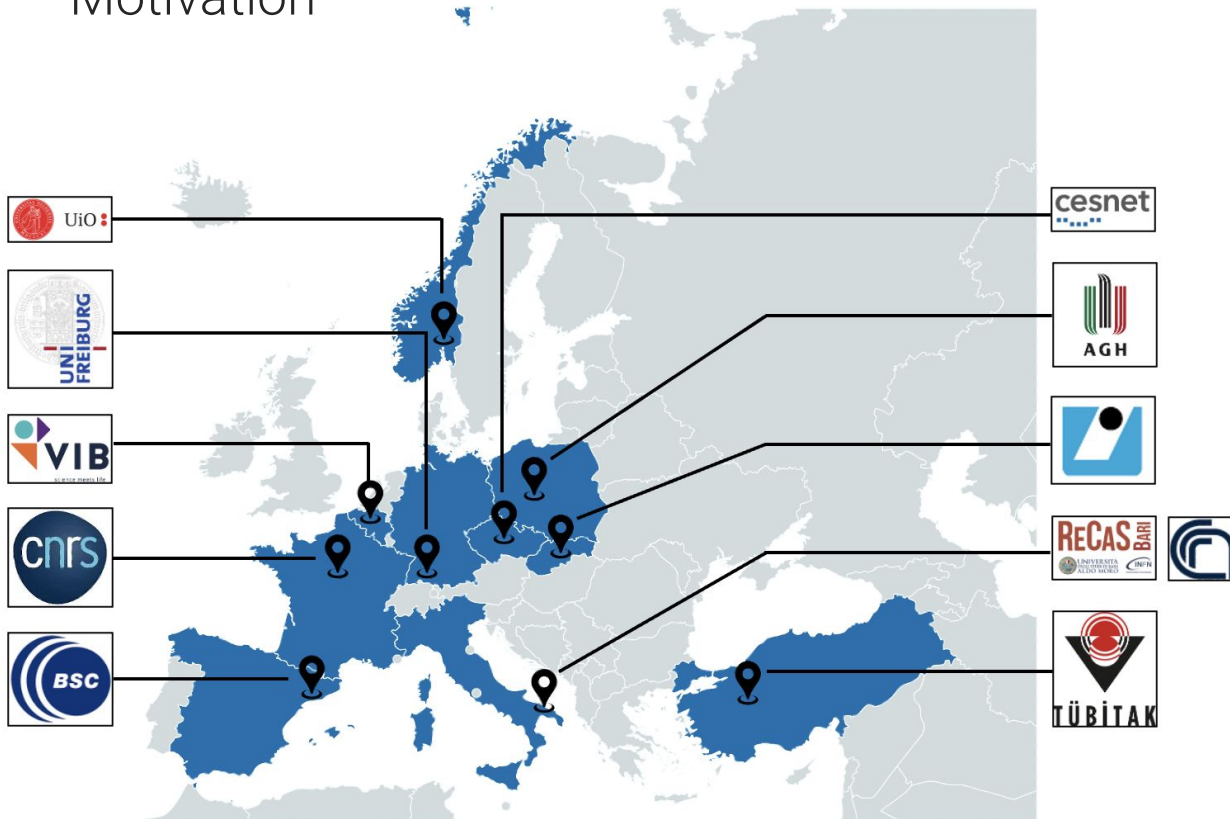
Overview

- The EuroScienceGateway project
- Broadening the login options for EuroScienceGateway
- BYOC: Bring Your Own Compute
- BYOS: Bring Your Own Storage
- Smart job scheduling across Europe



The EuroScienceGateway project

Motivation



National Cloud and HPC infrastructures have been established, with differences in

- Hardware
- Configuration
- Software stack
- Authentication and Authorization
- Access typically targeted at local researchers

goal: provide efficient and structured access to data, tools and workflows supported by suitable IT infrastructures.

The EuroScienceGateway project

How?

The screenshot displays the Galaxy Europe web interface. The top navigation bar includes 'Galaxy Europe', a home icon, and menu items for 'Workflow', 'Visualize', 'Data', 'Help', 'User', and a notification bell. A 'Using 0%' indicator is visible in the top right.

The left sidebar contains a navigation menu with categories: Upload, Tools, Workflows, Workflow Invocations, Visualization, Histories, History Multiview, Datasets, and Pages. The 'Tools' section is expanded, showing sub-categories: GENERAL TEXT TOOLS (Text Manipulation, Convert Formats, Filter and Sort, Join, Subtract and Group), GENOMIC FILE MANIPULATION (Convert Formats, FASTA/FASTQ, Quality Control, SAM/BAM, BED, VCF/BCF, Nanopore), COMMON GENOMICS TOOLS (Operate on Genomic Intervals, Fetch Sequences / Alignments), and GENOMICS ANALYSIS (Annotation, Multiple Alignments, Assembly, Mapping, Variant Calling).

The main content area features a search bar for tools, a 'Model in your automated workflows' section, and a central card titled 'The European Galaxy server'. This card contains a paragraph of text and several buttons: 'Browse installed tools', 'Request temporary increase of quota', 'Request TlaaS', 'Check the status of the server', and 'See the statistics'.

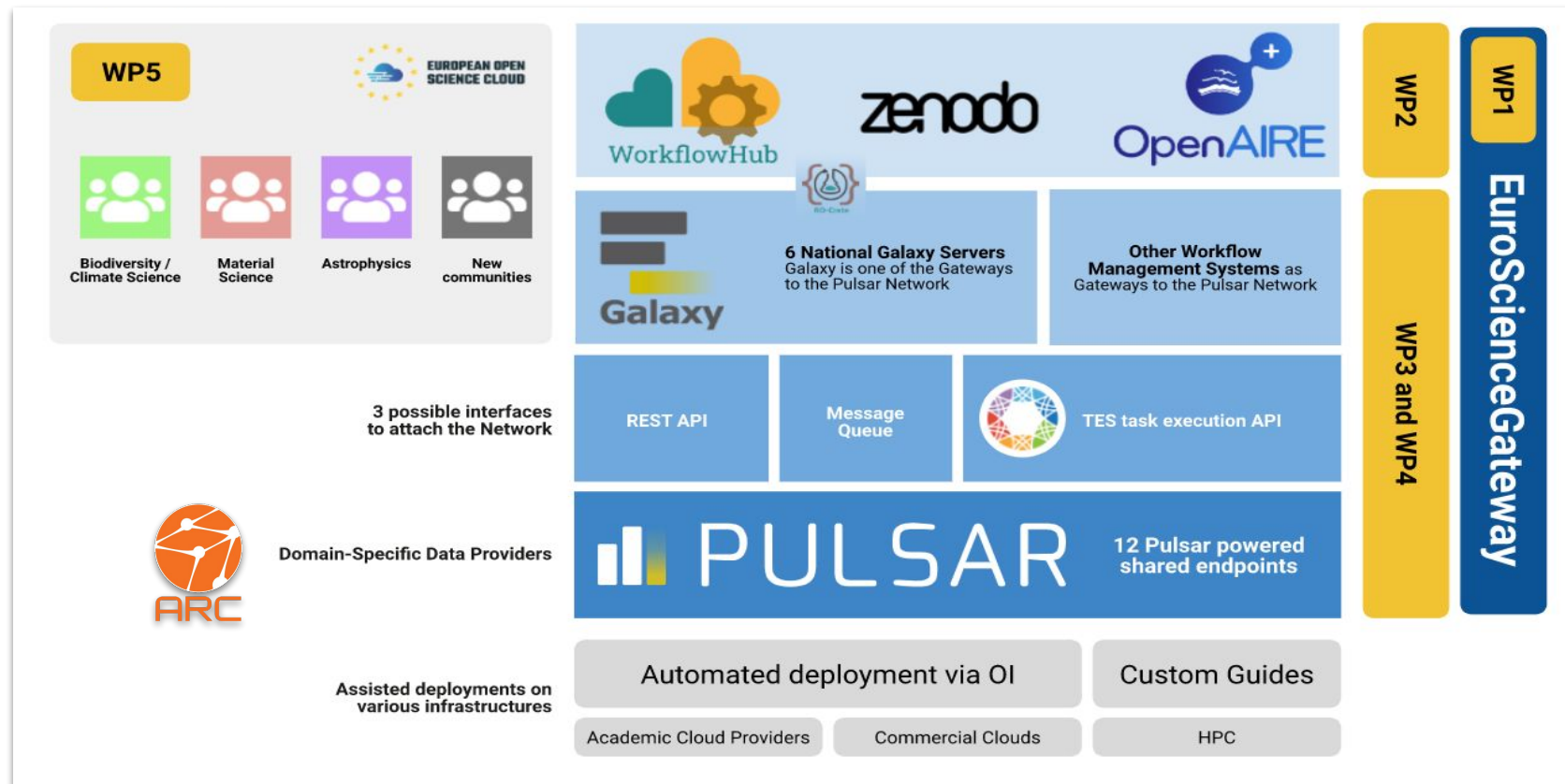
Below the server card are four informational cards: 'Projects', 'Communities', 'Citation', and 'Team', each with a brief description and a 'See all' button.

The right sidebar shows a 'History' section with a search bar for datasets and a message stating 'This history is empty. You can load your own data or get data from an external source.'

At the bottom of the main content area, there is a section titled 'Our Data Policy'.

The EuroScienceGateway project

How?



Partners



Overview

- The EuroScienceGateway project
- **Broadening the login options for EuroScienceGateway**
- BYOC: Bring Your Own Compute
- BYOS: Bring Your Own Storage
- Smart job scheduling across Europe



Broadening the login options for EuroScienceGateway

- Integration with WLCG IAM
- Integration with EGI Check-in



Broadening the login options for EuroScienceGateway

WLCG IAM: AAI solution for the Worldwide LHC Computing Grid

- WLCG IAM can be added to Galaxy thanks to [python-social-auth](#)
- Getting a token from WLCG IAM is one option in order to be allowed to submit jobs to ARC sites - IAM-services to trust are configurable in ARC

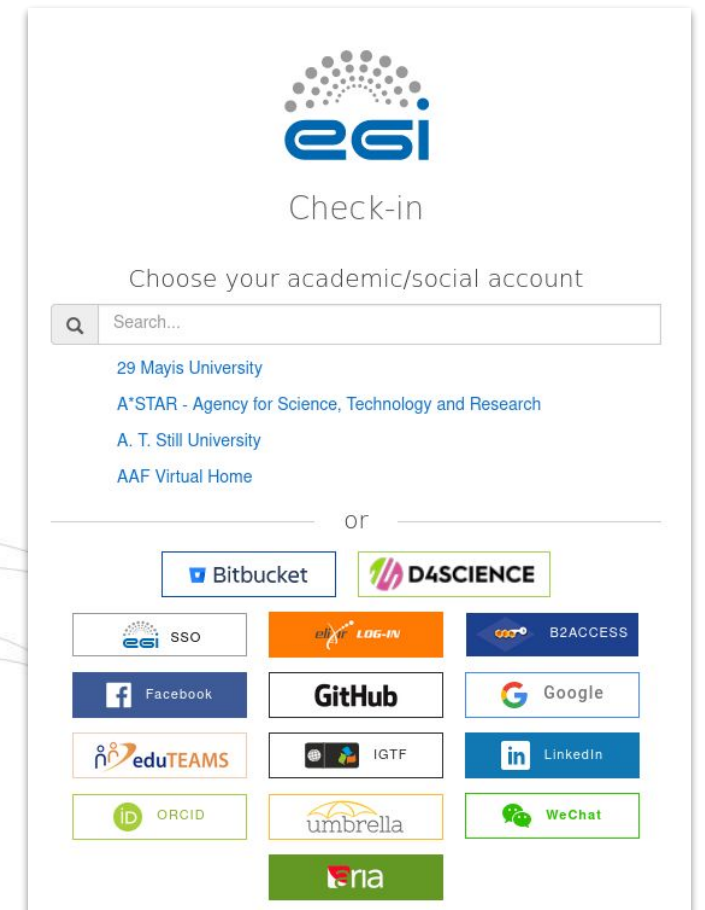
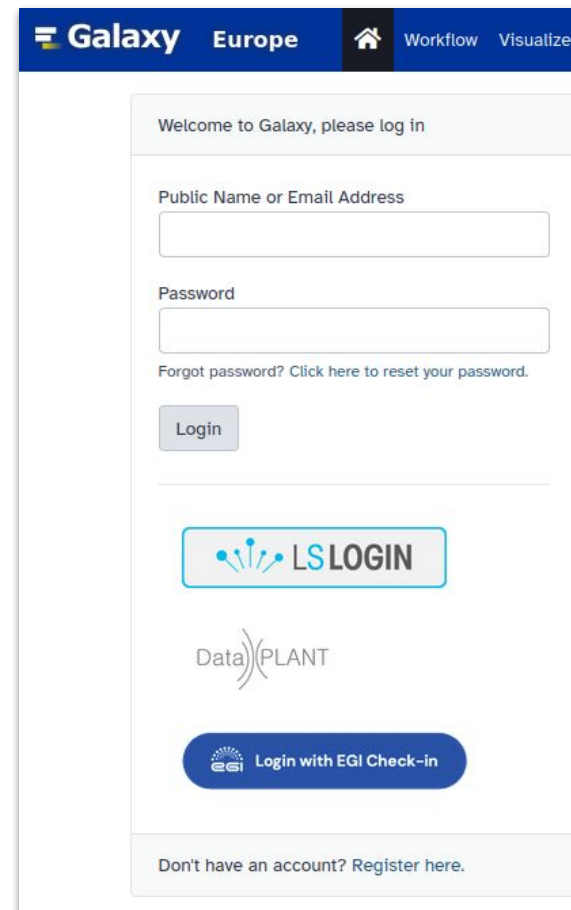
The screenshot shows the Galaxy web interface with a dark header containing the 'Galaxy' logo and navigation links for 'Workflow', 'Visualize', and 'Shared Data'. The main content area features a login form titled 'Welcome to Galaxy, please log in'. The form includes fields for 'Public Name or Email Address' and 'Password', a 'Forgot password? Click here to reset your password.' link, and a 'Login' button. At the bottom of the form, there is a WLCG logo and a link to 'Don't have an account? Register here.'

The screenshot shows the WLCG IAM login page at 'd.cnaf.infn.it/login'. It features the WLCG logo and the text 'Welcome to wlcg'. Below this, there are instructions to 'Sign in with your wlcg credentials' and a form with 'Username' and password fields, followed by a 'Sign in' button and a 'Forgot your password?' link. There are also options to 'Or sign in with' 'Your X.509 certificate' or 'CERN SSO'. A 'Not a member?' link leads to an 'Apply for an account' button. At the bottom, it states 'You have been successfully authenticated as CN=Maiken Pedersen maikemp@uio.no,O=Universitetet i Oslo,C=NO,DC=tcs,DC=terena,DC=org'.

Broadening the login options for EuroScienceGateway

EGI Check-in: AAI solution for the EGI Federated Cloud

- Use your existing credentials to log into Galaxy
- With a token from EGI Check-in you can deploy compute and storage resources in the EGI Federated Cloud and connect them with Galaxy (more details later)
 - Can also be used to submit jobs to ARC



Overview

- The EuroScienceGateway project
- Broadening the login options for EuroScienceGateway
- **BYOC: Bring Your Own Compute**
- BYOS: Bring Your Own Storage
- Smart job scheduling across Europe



Bring Your Own Compute

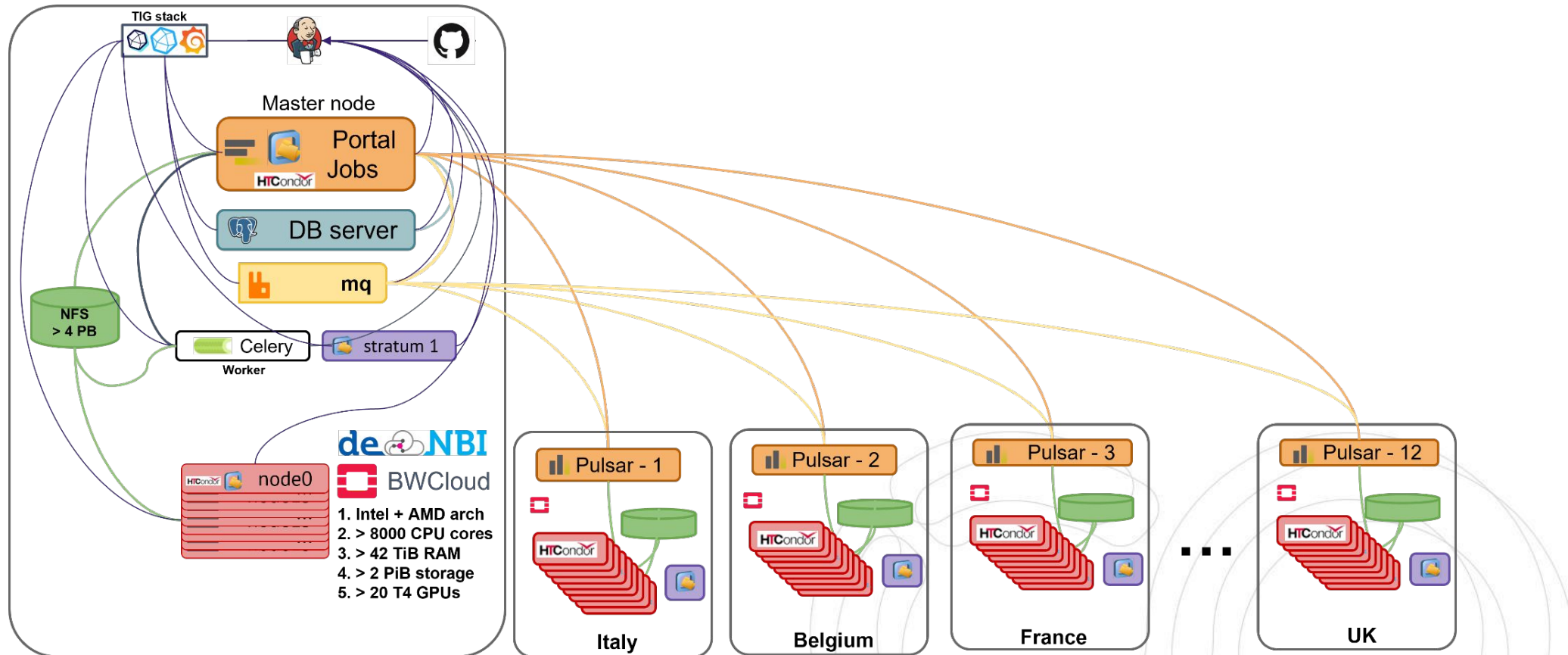
Connecting Galaxy with Pulsar

- Overview of Galaxy and Pulsar deployments in EU



Bring Your Own Compute

Connecting Galaxy with Pulsar



Bring Your Own Compute

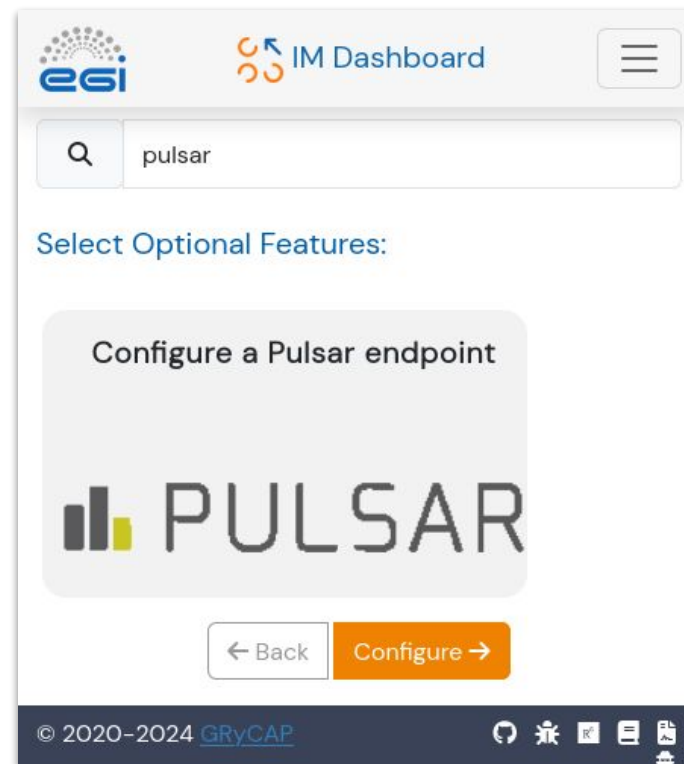
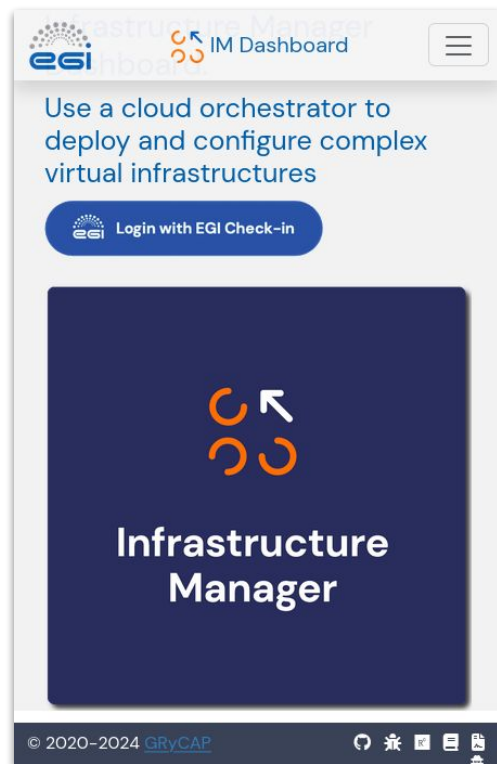
- Connect Galaxy with computing resources in the cloud
 - Connecting Galaxy with a managed ARC site
 - ARC deployment in the cloud with Infrastructure Manager
 - Pulsar deployment in the cloud with Infrastructure Manager
 - work by INFN to automate the connection of pulsar with Galaxy



Bring Your Own Compute

Pulsar deployment in the cloud with Infrastructure Manager

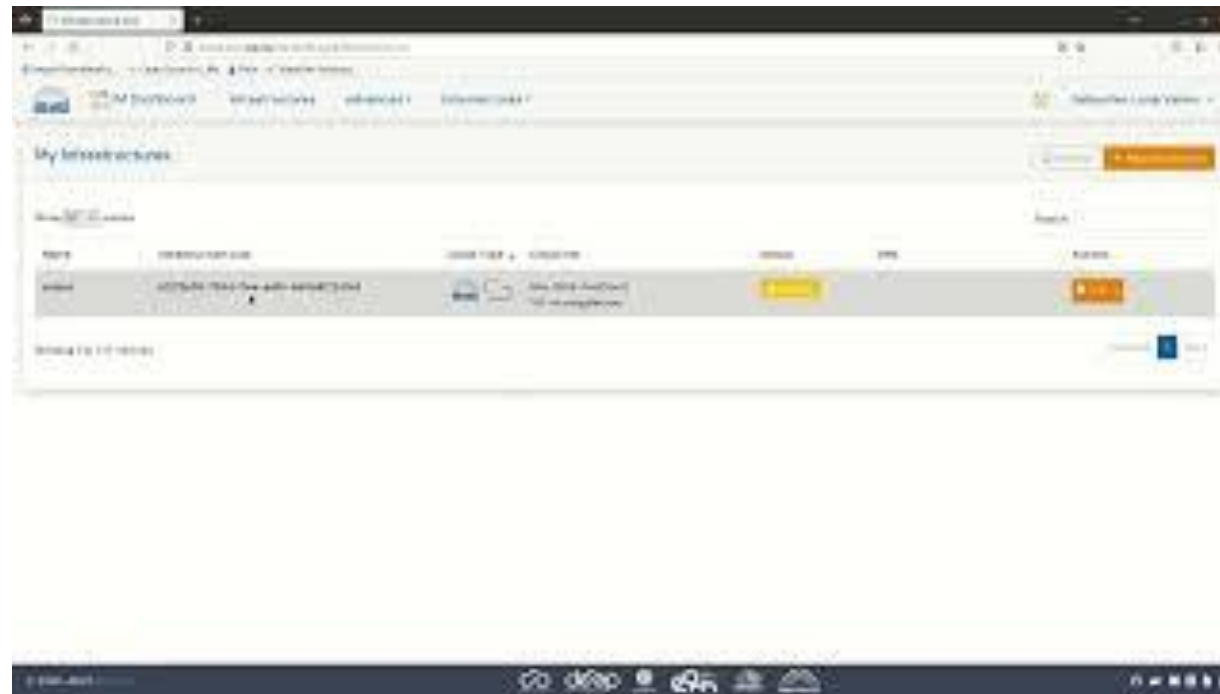
- Do have access to a cloud? Use Infrastructure Manager (<https://im.egi.eu/>) to deploy Pulsar and connect it to Galaxy



Bring Your Own Compute

Pulsar deployment in the cloud with Infrastructure Manager

- See tutorial in <https://galaxyproject.org/news/2023-10-31-esg-byoc-im/> and come to the booth to try it live!



Bring Your Own Compute

Connecting Galaxy with Pulsar

Bring your own Pulsar endpoint to Galaxy. You can add here your Pulsar credentials and specifications. After 24 hours Galaxy's job scheduling systems will take your Pulsar into account and schedule appropriate jobs to your compute resources. This is an experimental feature. Contact us if you want to learn more about it.

RabbitMQ custom username (e.g. mypulsarendpoint).

RabbitMQ custom alphanumeric password, at least 10 characters long (e.g. MyPulS4rP4ssw0rd).

Maximum number of CPU cores available.

Maximum number of RAM available (in GB).

Minimum number of GPU available (set this to zero if you don't want to share GPUs).

Maximum number of GPU available (set this to zero if you don't want to share GPUs).

Use distributed compute resources

Remote resources id

Select Value

- Brussel (Belgium) - VIB
- Prague (Czech Republic) - MetaCentrum**
- Bratislava (Slovakia) - IISAS
- Bari (Italy) - INFN
- Barcelona (Spain) - BSC-CNS
- Ankara (Turkey) - TÜBİTAK ULAKBİM
- Krakow (Poland) - Cyfronet
- Heraklion-Crete (Greece) - HCMR
- default - Galaxy will decide where to put your job**

Bring Your Own Compute

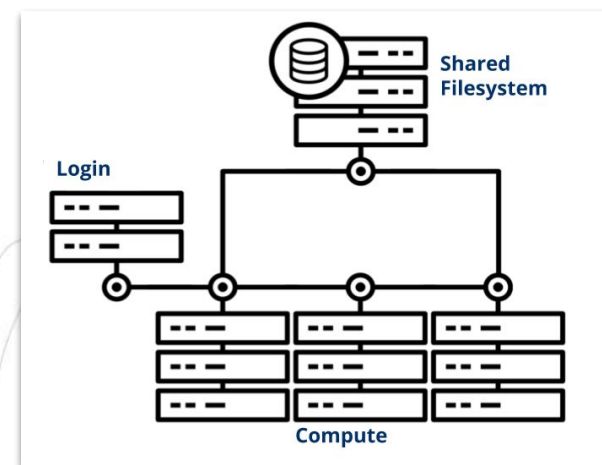
Connecting Galaxy with a managed ARC site - testing only, not yet in production

- Do you have access to a managed cluster with ARC? Log into Galaxy with WLCG IAM or EGI Check-in and configure the ARC endpoint in Galaxy

The URL to your ARC remote HPC compute resource. ^

Add your remote resource url (example: <https://arctestcluster-slurm-el9-arc7-ce1.cern-test.uiocloud.no>)

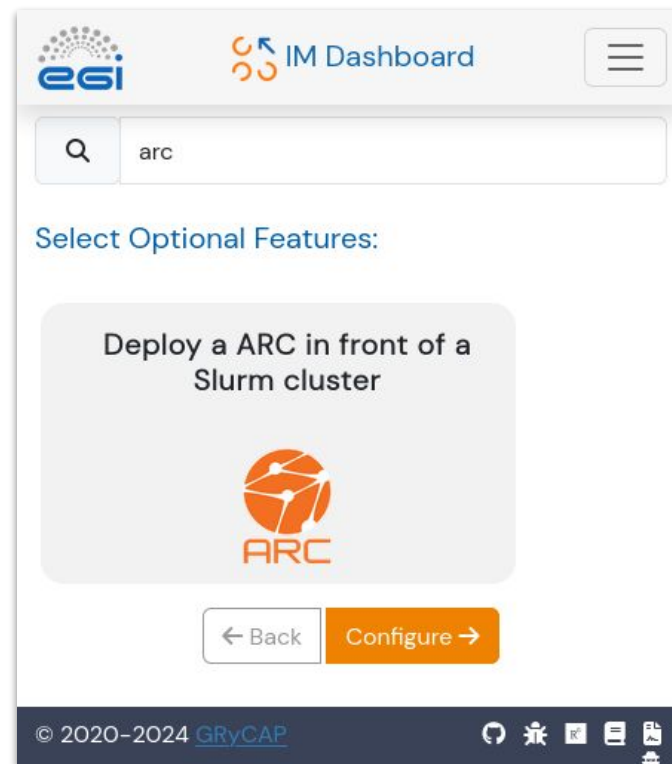
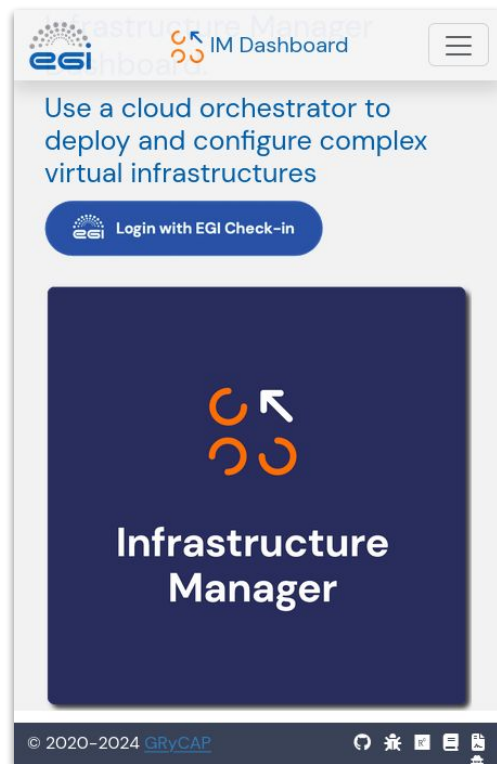
- Galaxy can execute jobs in ARC sites thanks to the prototype job runner [developed within the project](#)



Bring Your Own Compute

ARC deployment in the cloud with Infrastructure Manager

- Do have access to a cloud? Use Infrastructure Manager (<https://im.egi.eu/>) to deploy ARC and connect it to Galaxy



Overview

- The EuroScienceGateway project
- Broadening the login options for EuroScienceGateway
- BYOC: Bring Your Own Compute
- **BYOS: Bring Your Own Storage**
- Smart job scheduling across Europe



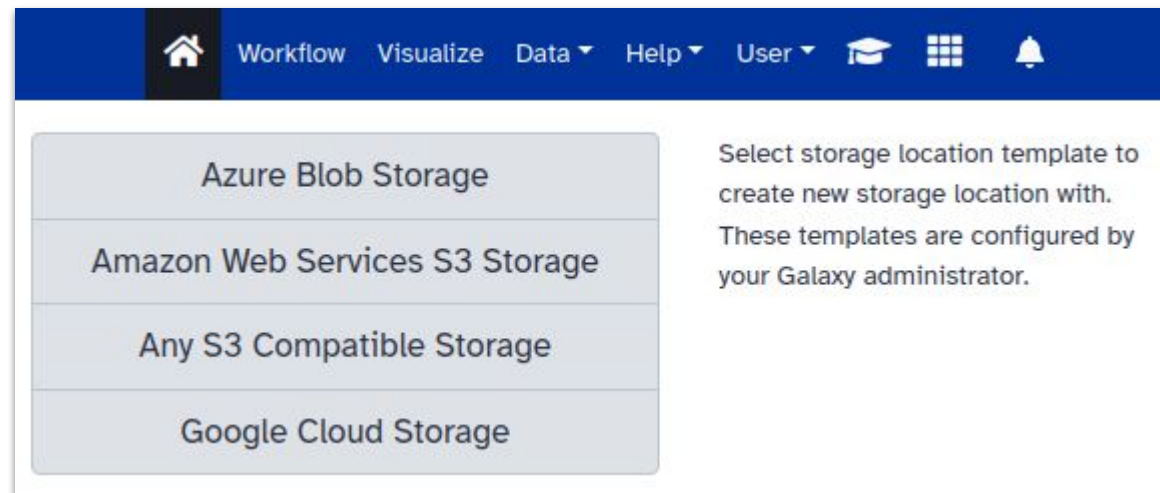
Bring Your Own Storage

Connect Galaxy with storage resources in the cloud

- Do you have access to object storage? Here is how to connect it to Galaxy
 - Tutorial: <https://galaxyproject.org/news/2024-09-20-esg-byos-im/>

- Select object storage type and provide
 - Name of the bucket / Container
 - Access (ID) key / Storage account
 - Secret key / Account key

- “Any S3 Compatible Storage” will also require
 - URL to the API endpoint



Bring Your Own Storage

Connect Galaxy with storage resources in the cloud

- Do you have access to object storage? Here is how to connect it to Galaxy
 - Tutorial: <https://galaxyproject.org/news/2024-09-20-esg-byos-im/>

- Select object storage type and provide
 - Name of the bucket / Container
 - Access (ID) key / Storage account
 - Secret key / Account key
- “Any S3 Compatible Storage” will also require
 - URL to the API endpoint

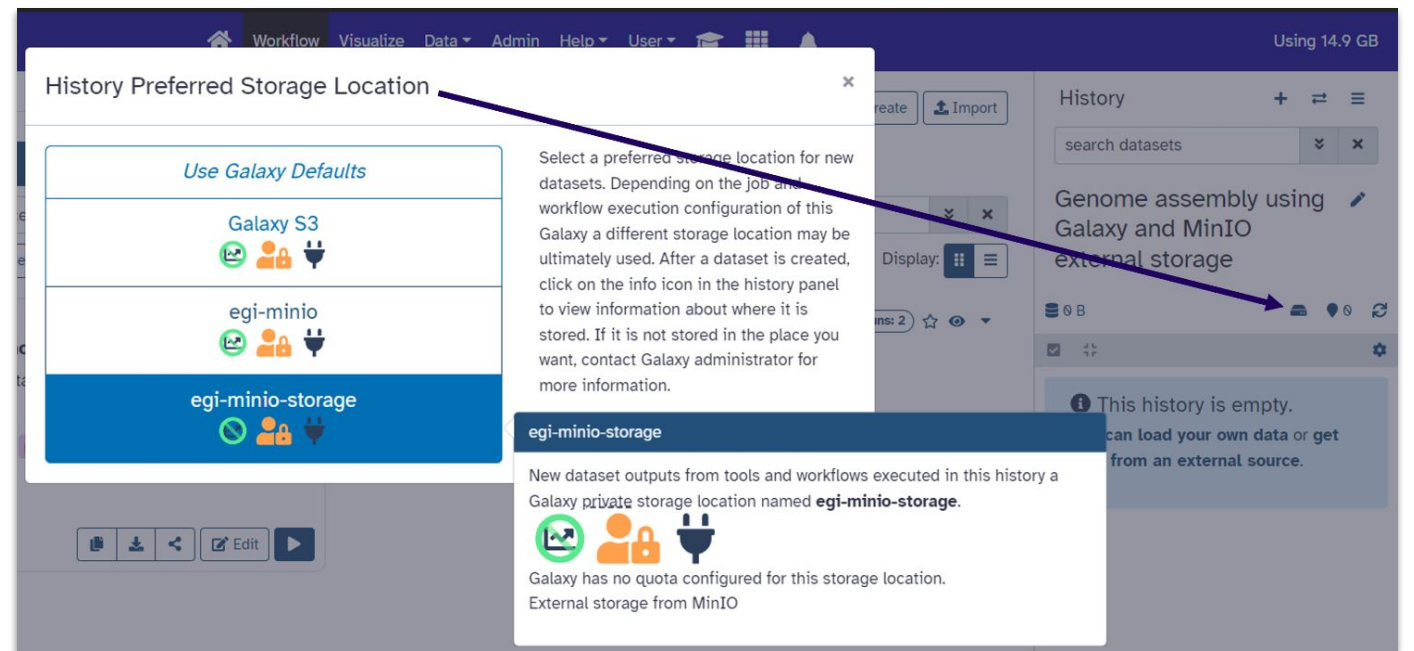
The screenshot shows the Galaxy Europe interface with a table of storage locations. A blue notification bar at the top says "Created storage location egi-minio-storage". The table has columns for Name, Description, Type, and From Template. The 'egi-minio-storage' row is highlighted, and a blue arrow points to it from the text "Your newly added external storage".

Name	Description	Type	From Template
Galaxy S3	My object store for all my data analysis needs in Galaxy	b01q3	Any S3 Compatible Storage
egi-minio		b01q3	Any S3 Compatible Storage
egi-minio-storage	External storage from MinIO	b01q3	Any S3 Compatible Storage

Bring Your Own Storage

Connect Galaxy with storage resources in the cloud

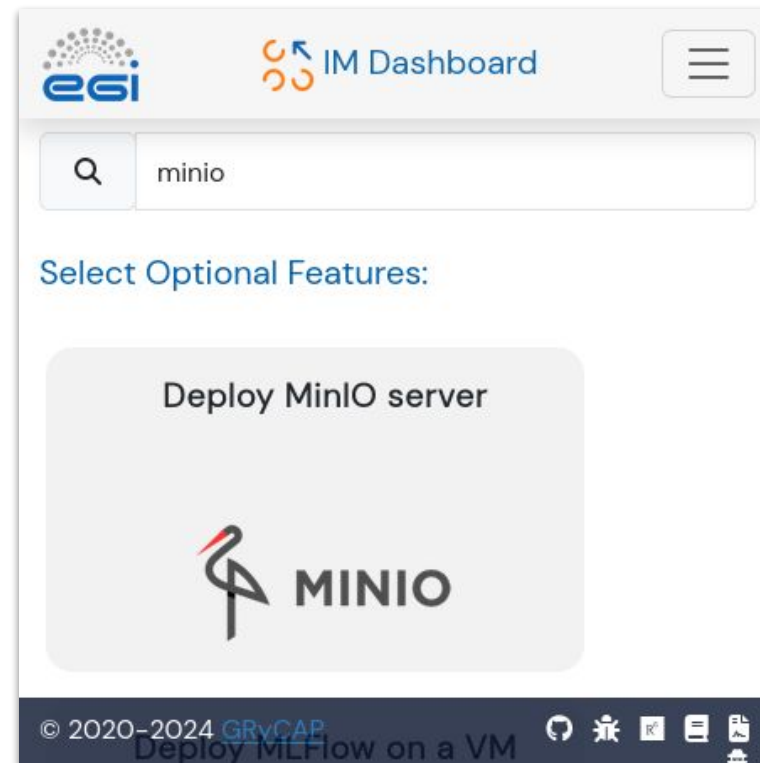
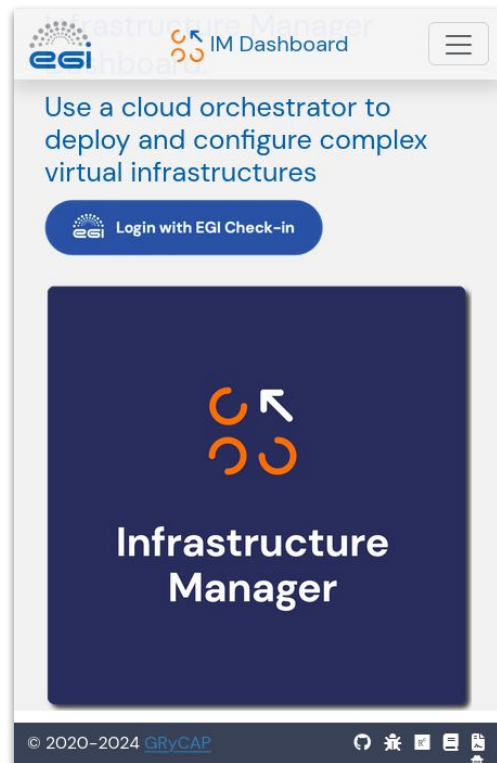
- Do you have access to object storage? Here is how to connect it to Galaxy and a tutorial is available [here](#).
- The new storage can be used
 - per History
 - per Tool
 - per Workflow
 - or Default (for everything)



Bring Your Own Storage

Connect Galaxy with storage resources in the EGI Federated Cloud

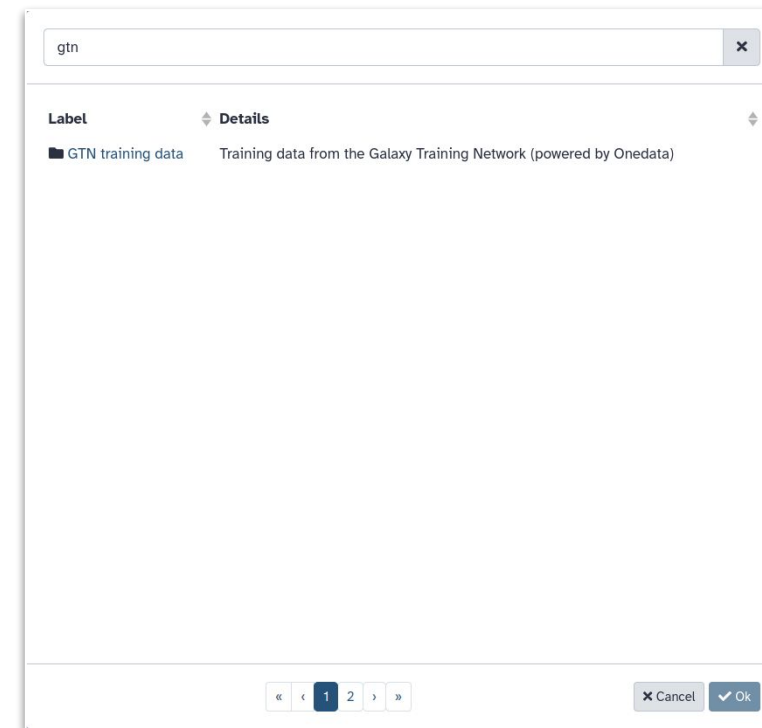
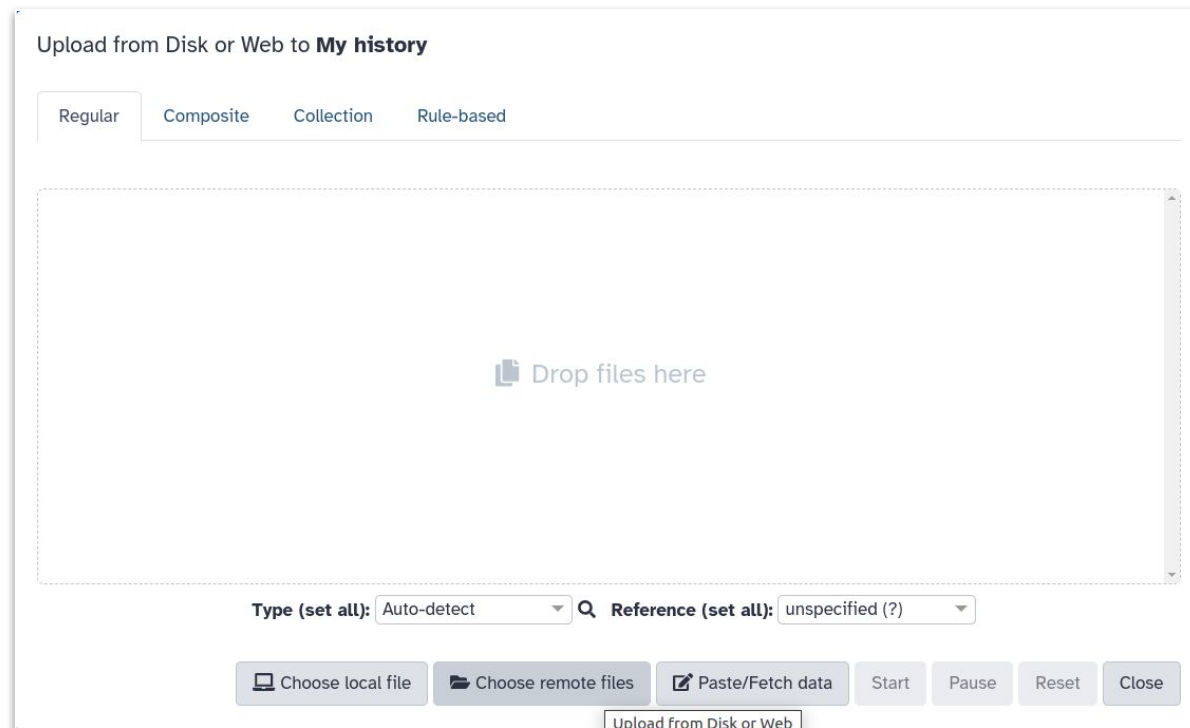
- Use Infrastructure Manager (<https://im.egi.eu/>) to deploy MinIO and connect it to Galaxy (see [tutorial](#))



Bring Your Own Storage

Connect Galaxy with storage resources in EGI DataHub

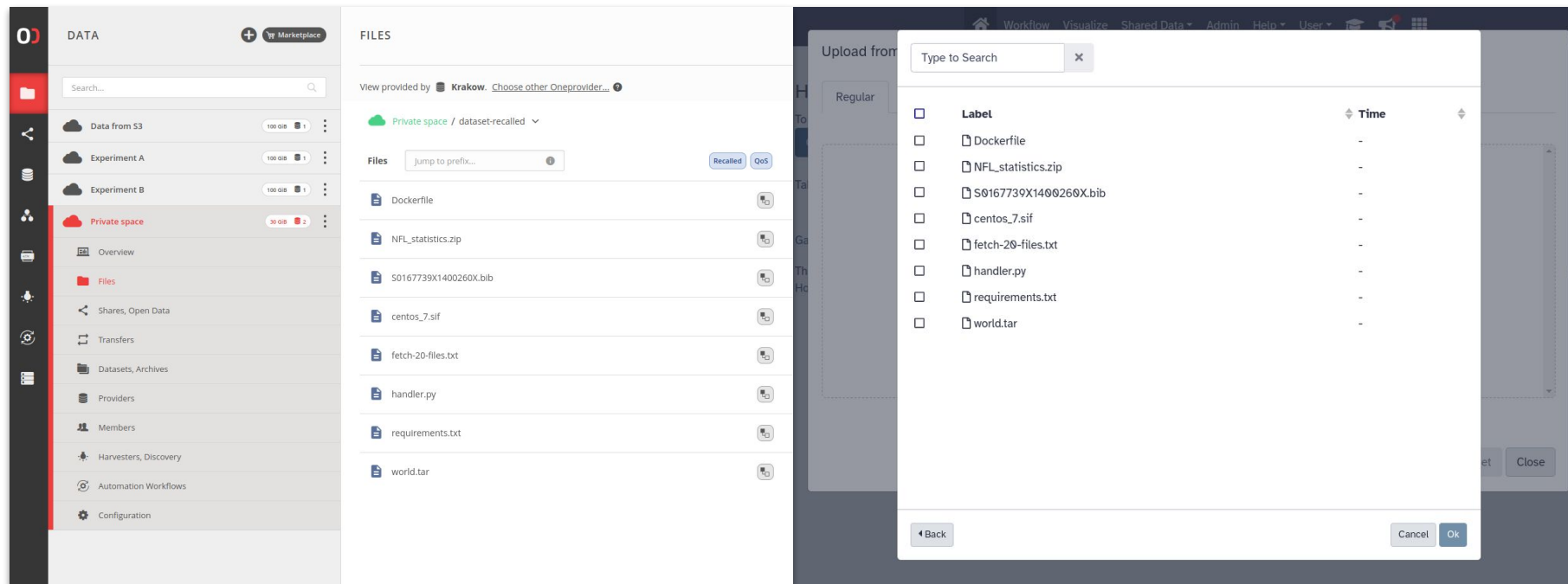
1. GTN training data is now hosted in EGI DataHub and available on *usegalaxy.eu* for anyone to import.



Bring Your Own Storage

Connect Galaxy with storage resources in EGI DataHub or *any Onedata environment*

2. Configure personal Onedata files source credentials to import or export your datasets.



Bring Your Own Storage

Connect Galaxy with storage resources in EGI DataHub *or any Onedata environment*

3. Configure a personal Onedata storage location for Galaxy user data (Object Store) – available in an upcoming release.

Create a new storage location for your data

Name *

Label this new storage location with a name.

Description - optional

Provide some notes to yourself about this storage location - perhaps to remind you how it is configured, where it stores the data, etc..

Onezone Domain

Domain of the Onezone service (e.g. datahub.egi.eu) to connect to.

Disable tls certificate validation?

No

Allows connection to Onedata servers that do not present trusted SSL certificates. SHOULD NOT be used unless you really know what you are doing.

Space Name

The name of the Onedata space where the Galaxy data will be stored. If there is more than one space with the same name, you can explicitly specify which one to select by using the format <space_name>@<space_id> (e.g. demo@7285220ecc636075ae5759aec7ad65d3cha8f9).

Galaxy root directory

The relative directory path in the space at which the Galaxy data will be stored. If empty, the data will be stored in the space's root directory.

Access Token

Your access token, suitable for REST API access in a Oneprovider service. Must allow both read and write data access.

Overview

- The EuroScienceGateway project
- Broadening the login options for EuroScienceGateway
- BYOC: Bring Your Own Compute
- BYOS: Bring Your Own Storage
- **Smart job scheduling across Europe**



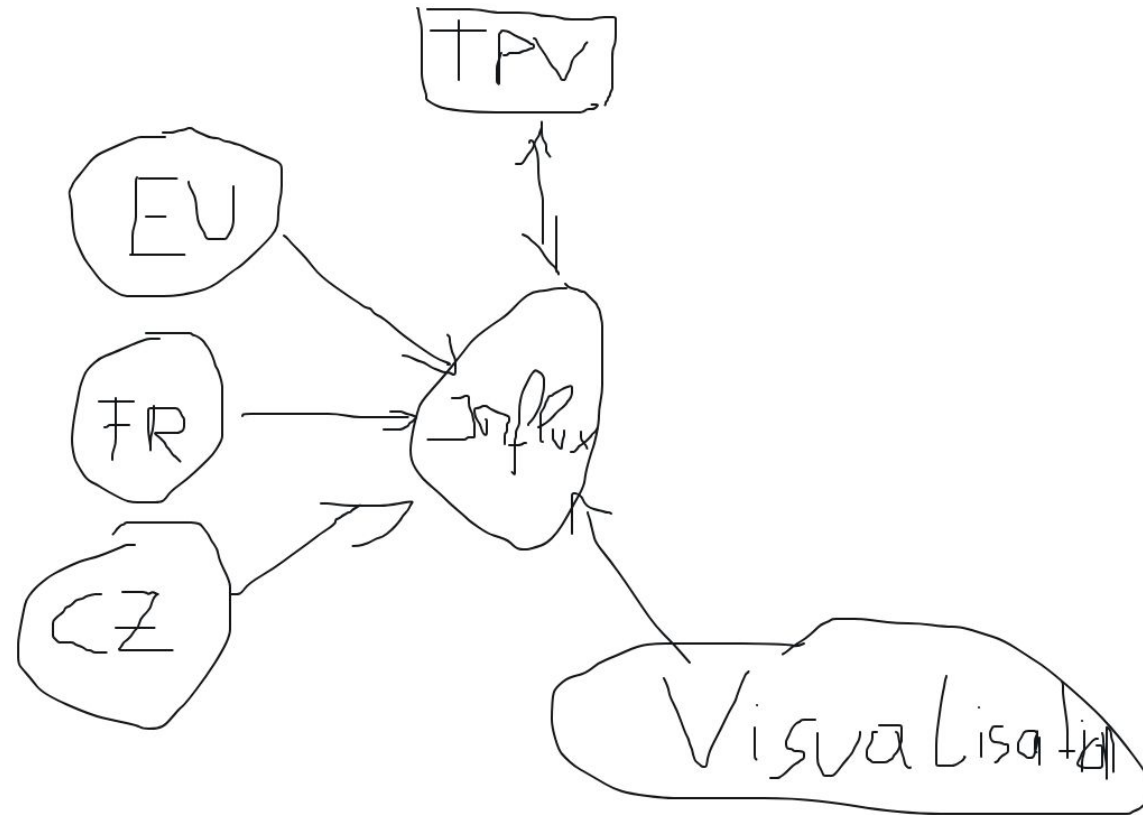
Smart scheduling

- How do we efficiently schedule jobs?
- Gather usage statistics from pulsar endpoints
- TPV, metascheduling algorithm, and TPV metascheduler API
- Visualization of jobs across Europe



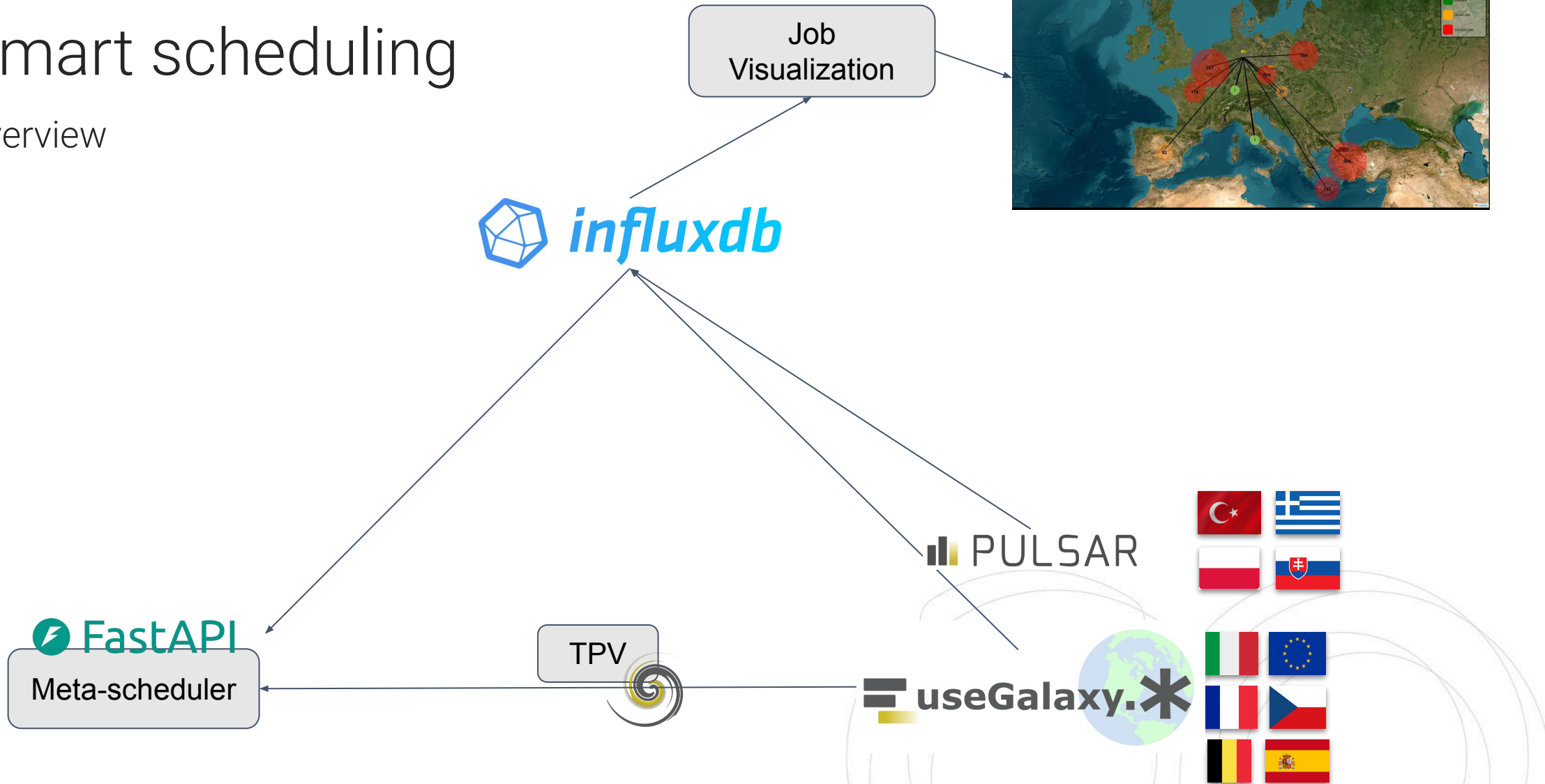
Smart scheduling

Brainstorming over Zoom :)



Smart scheduling

Overview



Smart scheduling

Central API endpoint

- Implements **matchmaking** logic, ranking available pulsar destinations based on:
 - current - and historically collected **(tool-destination) metrics**
 - **data locality** (geolocation) of the objectstore/destination
- Currently, simple weighting of the different metrics
- Working on more advanced Fuzzy/Adaptive-based matchmaking algorithms

```

1  tools:
2    default:
3      ...
4      rank: |
5        ...
6
7        # currently object store info is stored in a yaml
8        request_data["static_objectstores_info"] = objectstore_info
9
10       # dataset info
11       request_data["static_dataset_info"] = helpers.get_dataset_attributes(job.input_datasets)
12
13       # static job info
14       request_data["static_job_info"] = {"tool_id": tool.id, "mem": mem, "cores": cores, "gpu": gpu,}
15
16       # current destination info
17       dest_info = []
18       for dest in candidate_destinations:
19         dest_dict = {"id": dest.id}
20         dest_dict.update(dest.context)
21
22         dest_dict["queued_job_count"] = model.Query(model.Job).filter(
23             model.Job.state == "queued",
24             model.Job.destination_id == dest.dest_name).count()
25         dest_dict["running_job_count"] = model.Query(model.Job).filter(
26             model.Job.state == "running",
27             model.Job.destination_id == dest.dest_name).count()
28
29         dest_info.append(dest_dict)
30
31       request_data["current_dest_info"] = dest_info
32
33       # Send a POST request to the API endpoint
34       response = requests.post(api_url, json=request_data)
35
36       sorted_candidate_destinations = sorted(candidate_destinations,
37                                               key=lambda x: sorted_destination_ids.index(x.id))
38       sorted_candidate_destinations
39

```

Smart scheduling

Central API endpoint

- Implements **matchmaking** logic, ranking available pulsar destinations based on:
 - current - and historically collected **(tool-destination) metrics**
 - **data locality** (geolocation) of the objectstore/destination
- Currently, simple weighting of the different metrics
- Working on more advanced Fuzzy/Adaptive-based matchmaking algorithms

POST /process_destinations Process Destinations

Parameters Cancel Reset

No parameters

Request body **required** application/json

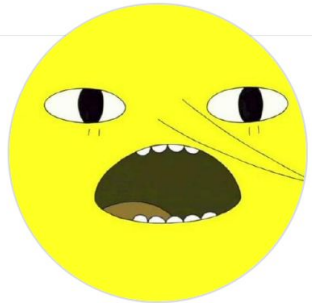
```

{
  "static_objectstores_info": {
    "object_store_italy": {"latitude": -26.4390917, "longitude": 133.281323}
  },
  "static_dataset_info": {
    0: {"object_store_id": "object_store_italy", "size": 1073741824000.0}
  },
  "static_job_info": {"tool_id": "trinity", "mem": 8, "cores": 2, "gpus": 0},
  "current_dest_info": [
    {
      "id": "slurm_germany",
      "latitude": 51.1642292,
      "longitude": 10.4541192,
      "queued_job_count": 5,
      "running_job_count": 5
    },
    {
      "id": "pulsar_italy",
      "latitude": 50.0689816,
      "longitude": 19.9070188,
      "queued_job_count": 5,
      "running_job_count": 5
    }
  ]
}
    
```

Execute Clear

Thanks!

All ESG members, specially:



Paul De Geest



Sanjay Kumar Srikakulam



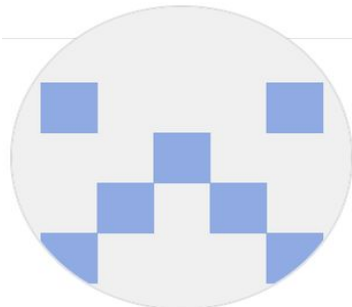
Abdulrahman Azab



Björn Grüning



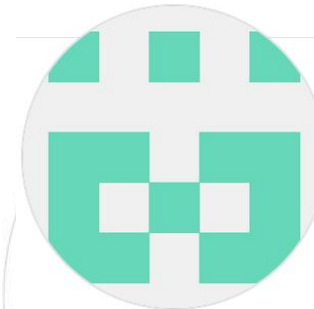
Enol Fernández



Maiken Pedersen



Łukasz Opiola



Martin Demko

Tomáš Vondrák